12-1-2007

Improved Covariance Model Parameter Estimation Using RNA Thermodynamic Properties

Jennifer A. Smith
Boise State University

Kay C. Wiese
Simon Fraser University

This document was originally published by IEEE in the *Bio-Inspired Models of Network, Information and Computing Systems, 2007. Bionetics 2007*. Copyright restrictions may apply. DOI: 10.1109/BIMNICS.2007.4610108

Improved Covariance Model Parameter Estimation Using RNA Thermodynamic Properties

Scott F. Smith
Boise State University
ECE Department
Boise, Idaho, 83725-2075, USA
+1-208-426-5743
sfsmith@boisestate.edu

Kay C. Wiese
Simon Fraser University
School of Computing Science
Surrey, BC, Canada, V3T 0A3
+1-778-782-7436
kwiese@cs.sfu.ca

ABSTRACT

Covariance models are a powerful description of non-coding RNA (ncRNA) families that can be used to search nucleotide databases for new members of these ncRNA families. Currently, estimation of the parameters of a covariance model (state transition and emission scores) is based only on the observed frequencies of mutations, insertions, and deletions in known ncRNA sequences. For families with very few known members, this can result in rather uninformative models where the consensus sequence has a good score and most deviations from consensus have a fairly uniform poor score. It is proposed here to combine the traditional observed-frequency information with known information about free energy changes in RNA helix formation and loop length changes. More thermodynamically probable deviations from the consensus sequence will then be favored in database search. The thermodynamic information may be incorporated into the models as informative priors that depend on neighboring consensus nucleotides and on loop lengths.

Keywords

Bioinformatics, Covariance models, RNA secondary structure, Database search, Non-coding RNA gene search.

1. INTRODUCTION

Covariance models (CMs) have a well-established history of use in non-coding RNA (ncRNA) gene search [1, 2]. They can be thought of as an extension to profile hidden Markov models [3] that take into account RNA secondary structure in addition to primary sequence. The additional information about which model positions are expected to base pair with which other model positions is needed in ncRNA gene search since there is selection pressure to maintain the shape of ncRNA molecules. Simultaneous mutation of two model positions that are widely separated in sequence such that one Watson-Crick pair is substituted for another may not impede molecular function, whereas a single mutation in the pair might be fatal. Hence, an estimated probability distribution over the sixteen possible pairs of nucleotides at the two positions (a joint distribution) can be much more useful than two individual (marginal) distributions over four possible nucleotides each.

Estimation of covariance model parameters starts with the formation of a multiple alignment of known members of a ncRNA family annotated with secondary structure (base pairing) information. The model takes the form of a tree of states with probability distributions over symbol (nucleotide) emissions by states and transitions between states. Emission probabilities are estimated from observed frequencies of the four nucleotides (A, C, G, and U) in unpaired alignment columns and of nucleotide pairs in paired columns. Transition probabilities are estimated from observed position-dependent frequencies of insertions and deletions in the multiple alignment.

More than half of the 594 covariance models available in the Rfam ncRNA database [4] are estimated using six or fewer known family members (Rfam 8.0, updated February 2007 [5]). As a result, most emission and transition probabilities are based on no observed occurrences of the emission or transition event. The simplest way to handle this is to add pseudocounts to all events such that no probability is estimated as identically zero. This amounts to applying a prior to the estimator with the information that no possibility should be completely ruled out. As the number of sequences used to estimate the parameters increases, the preponderance of evidence simply makes these non-zero probabilities very small relative to observed events.

The most often used package for CM parameter estimation and CM-based ncRNA gene search (including the formation of the Rfam database) is Infernal [6]. In November 2005 this package began to incorporate Dirichlet mixture priors into the parameter estimation algorithm (with version 0.6). This allows for much more informative prior information based on observed frequencies of mutation, insertion, and deletion events in ncRNA molecules in general (not just in the specific family members used to estimate the model). This is a major step forward in that CM parameter files now incorporate a bias towards accepting AU, UA, GC, CG, GU, and UG even if these pairs were never observed in the multiple alignment data. Furthermore, insertion and deletion penalties now depend on whether they occur in a helix or a loop even if no insertions or deletions are observed in the multiple alignment.

In this work we propose to go one step further in supplying prior information by incorporating measured thermodynamic properties of RNA [7]. This information is already widely used in algorithms that estimate RNA secondary structure from sequence [8, 9]. The

Bionetics'07, December 10–12, 2007, Budapest, Hungary.
Copyright 2007 ICST 978-963-9799-11-0.

thermodynamic measurements give another independent source of information for emission and transition probabilities. More importantly, the priors generated depend on the consensus sequence and secondary structure of the ncRNA family. As will be seen in the Parameter Estimation section below (sec. 3), loop lengths and neighboring base pairs in helices make a difference in the likelihood of insertions, deletions, and paired base emissions.

Good CM parameter estimation is important not only for high quality of ncRNA gene search results, but also for the speed of modern CM-based search algorithms. In the Query-Dependent Banding method [10] used since version 0.71 of Infernal (November 2006), the state-dependent bounds on subsequence length are calculated using the parameters of the specific family being searched. The potential to improve search performance through limiting the subsequence length component of the search space has also been noted in [11]. As can be seen in the timing data of Weinberg and Ruzzo [12], CM-based ncRNA search can take a single processor years to search for members of a single family (and each family needs to be done independently), so performance is very important.

The remainder of this paper is structured as follows. Section 2 gives an overview of the organization of covariance models and their use in ncRNA gene search. The estimation of CM parameters using both traditional observed frequency approaches and the proposed thermodynamics approach is presented in Section 3. Concluding remarks may be found in Section 4.

2. COVARIANCE MODELS

The structure of a covariance model is determined by the consensus secondary structure of the ncRNA family multiple alignment. Pseudoknots [13] in the secondary structure are not allowed in a CM. If pseudoknots occur in the real structure, some base-paired positions must be represented as if they were unpaired (which results in some loss of model explanatory power). As a result, only three classes of consensus alignment columns exist: singlets, left pairs, and right pairs. Singlets are associated with left-emitting (L) or right emitting (R) nodes in the CM and each pair of columns is associated with a single pair-emitting (P) node. The other non-emitting types of nodes, start (S), end (E), and bifurcation (B) are used to organize the nodes into a tree.

At the top of the tree is a special S node (called the root start node) and the bottom of the tree has one or more E nodes. Each node represents a sub-model covering a contiguous range of the consensus columns of the multiple alignment. The E nodes represent a null model (with a specific sequence position, but of length zero). Emitting nodes build on the representation of their child node by adding a position to the right (R), left (L), or both (P). B nodes have two child nodes (all others have one child) and merge two contiguous sub-models into a single contiguous model. The two B-node children are always S nodes such that each of the B-node children could be interpreted as a complete CM in their own right.

In order to allow for insertions and deletions, every CM node type has a particular internal state structure. R-type, L-type, and P-type states are symbol-emitting just as the R, L, and P nodes are. Symbols might be emitted representing the consensus function of the node, in which case the R-type, L-type, and P-type nodes are designated MR, ML, and MP respectively (where M is for match). Insertions are allowed through the use of additional R- and L-type states designated IR and IL respectively. D states are non-emitting and allow the node to be bypassed. B, S, and E nodes contain B, S, and E states respectively representing their consensus function. The node tree completely determines the associated state tree, since the internal state arrangement of each node type is fixed. The state tree is what is actually processed when doing a database search.

An example multiple alignment is shown in Figure 1 for the Cardiovirus cis-acting replication element (CRE) ncRNA family from Rfam (designated RF00453). Only the twelve sequences used to estimate the original model (the seed sequences) are shown, but a further 18 sequences have been found using the estimated CM through database search. This family happens to have no observed insertions or deletions, so all alignment columns are consensus columns. The next-to-last row in the alignment is the consensus secondary structure, where "<" and ">" are left and right half of a base pair respectively and "." is used for a singlet column. The final row shows the consensus sequence.

The Cardiovirus CRE covariance model has a single E node located between the leftmost ">" and the "." to its left. Ten L nodes represent the loop at the end of the hairpin, each representing an additional "." to the left of the string of dots represented by its child node. Three P nodes represent three "<" and ">" pairs working outward from the core of eleven singlets. A bulge loop on the 5' side of the helix is represented by three L nodes (group of three "." in Figure 1). Seven P nodes continue the helix. Finally, the root start node caps the node tree at the top. Since there are no bifurcations in this model, the node "tree" is a linear structure with no branching. The observed frequencies of symbol emissions are obtained by counting the number of A, C, G, and U in each column or AA, AC, ... , UU in each pair of columns. The observed frequencies of insertions between consensus columns or deletions of consensus columns are all zero for the sequences used to estimate this family model.

```
ACGGUCACAAACACCCAGUCAACAGUGGGCCGU
ACGGUCACAAACACCCAAUCAACCGUUGGUCGU
UCGGCCACAAACACACAGUCUACUGUUGGCCGG
UCGGCCACAAACACACAAUCUACUGUUGGUCGA
UCGGCCACAAACACACAAUCUACUGUUGGUCGG
UCGGCCGUAAACACCCAAUCAUCAGUAGGCCGA
ACGGUCACAAACACCCAAUCAACCGUGGGCCGU
UCGGCCACAAACACGCAGUCUACUGUUGGCCGA
UCGGCCACAAACACUCAAUCCUCCGUUGGCCGG
UCGGCCACAAACACACAAUCUACCGUUGGUCGA
UCGGCCACAAACACCCAAUCUACUGUUGGUCGA
ACGGUCACAAACACCCAAUCAACAGUAGGCCGU
<<<<<<<...<<<..........>>>>>>>>>>
UCGGCCACAAACACCCAAUCUACUGUUGGCCGA
```

Figure 1. Cardiovirus CRE (RF00453) multiple alignment

Figure 2 shows a graphical view of the consensus Cardiovirus CRE sequence and secondary structure. The letters show the most commonly occurring symbol at each consensus position, where the U at top left is on the 5'-end (sequence start end). The 5' bulge loop with consensus sequence CAA can be seen on the left.

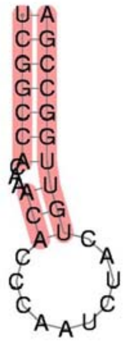

Figure 2. Cardiovirus CRE (RF00453) consensus structure

The covariance model tree for Cardiovirus CRE is shown in Figure 3. Since this ncRNA family has no bifurcations, the 'tree' has only a single branch. The model is evaluated starting at the E node (bottom right) and progressing up the tree toward the root start node (upper left). The line connecting the right and left half in the figure has no special meaning and it there only to allow the figure to fit better on the page.

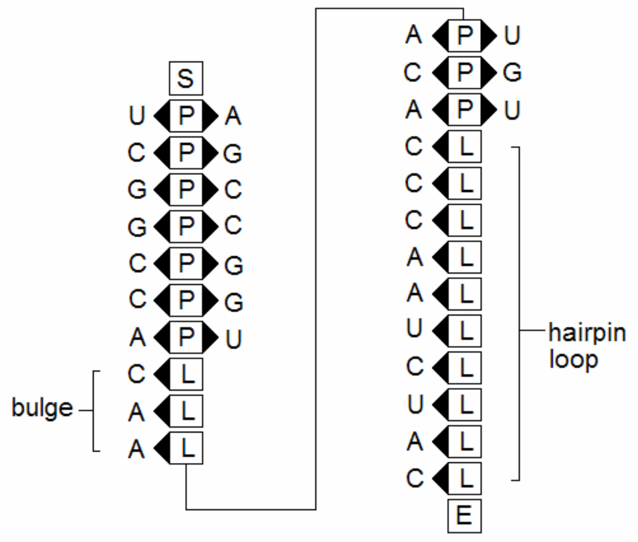

Figure 3. CM node 'tree' for Cardiovirus CRE (RF00453)

3. PARAMETER ESTIMATION

3.1 Observed Frequencies Within a Family

Given a set of aligned and structure-annotated RNA sequences, parameters for a CM model of the set of sequences can be found by counting the number of occurrences of each state transition and emission in the observed data. The transition and emission scores can then be set equal to the base 2 logarithm of the ratios of the counts to the number of samples. The resulting scores have units of bits such that a score increase of 1 implies a doubling of the likelihood that the database sequence fits the model.

Many refinements can be made to this basic counting process. Emission probabilities can be corrected for bias in the background frequencies of the various symbols. This might be important if the ncRNA family only occurs in a particular organism and the genome of that organism is either AU or GC rich. Entropy weighting can be applied to the sequences to correct for the fact that the sequence set may contain mostly very similar sequences and very few dissimilar sequences. This can occur when a well-studied model organism is overrepresented in the sequence set used to estimate the model.

As discussed briefly in the introduction, it is unusual to include absolutely no prior information, either implicitly or explicitly, in the model estimate. At the very least, pseudocounts are added to every possible event (allowed state transitions, symbol emissions, or symbol-pair emissions) such that no allowed transition or emission score is set to minus infinity. This represents prior information in the sense that the model builder is saying that the event should be allowed even though it has never been observed. More informative priors are discussed in the following section.

3.2 Dirichlet Mixture Priors for Observed Frequencies Outside a Family

The current version of Infernal allows for the specification of Dirichlet mixture [14] priors. A Dirichlet mixture is a weighted combination of Dirichlet priors. One advantage to the Dirichlet prior is that the posterior distribution has the same form as the prior distribution when used in a Bayesian parameter estimation setting, which results in simpler calculation. The details of using Dirichlet priors will not be presented here, but the fact that there is a mechanism to include information from non-family sequences is important. If simulated sequences with deviations from the family consensus can be generated with deviation frequencies proportional to those expected from thermodynamic measurement, then the existing software can turn these simulated sequences into an appropriate prior for use by the Infernal search program.

The existing use of the Dirichlet mixture prior in ncRNA search involves using a filtered database of real ncRNA sequences to build the prior. Most of these sequences will be from outside the particular family for which the prior will subsequently be used to build a model. As such, the prior is meant to represent generic features of all ncRNA molecules. Information such as the fact that Watson-Crick base pairs tend to substitute for one another more often than random pairs is contained in these priors. Other information includes the fact that insertions and deletions are less common in P nodes (helices) than in L or R nodes (loops).

The priors in current use are specified over singlet emissions or pair emissions independent of the surrounding consensus symbols. As will be seen in the next section, this is a drawback, since thermodynamic measurement indicates that neighboring bases are important. The current transition priors are specified separately for the different types of states and containing nodes on each end of the transition (73 combinations in all). However, a transition to a D state in the middle of a hairpin loop (formed from a sequence of L nodes) will be the same no matter what the overall length of the loop. It is shown in Section 3.4 that thermodynamic measurement indicates that this prior ought to be adjusted depending on consensus loop length.

Since the current Infernal package can not handle loop-length and neighboring nucleotide effects, the full extent of thermodynamic measurement information can not be incorporated into CM parameter estimation without altering the parameter estimation program (the Infernal package cmbuild program). However, a first step towards determining if the thermodynamic measurement information might be useful could be determined by seeing if a prior with similar effectiveness to the currently used prior can be built using the thermodynamic measurements in place of the general database sequences currently used.

3.3 Dependence of Free Energy Changes on Neighboring Bases

The free energy change as a result of a single-stranded RNA molecule forming hydrogen bonds with itself to generate a helix structure can be expressed as sum of terms [8]. There is a term for the helix as a whole which will be ignored in what follows since we will assume that the helix exists with or without a possible mutation if the RNA is in fact a true family member. There is a correction to the whole helix term that depends on whether the helix base sequences are symmetric which will also be ignored here, but which could be incorporated without too much trouble. Finally, there are a collection of terms which depend on the nucleotide composition of neighboring base pairs. These terms are sometimes called stacking energies or helix propagation energies and will be the focus of the discussion in this section.

Table 1. Free energy changes for helix propagation

Neighboring Bases	ΔG (kcal/mol)	Neighboring Bases	ΔG (kcal/mol)
5'-AA-3' 3'-UU-5'	-0.9	CC GG	-2.9
AU UA	-0.9	CG GC	-2.0
UA AU	-1.1	GC CG	-3.4
AC UG	-2.1	CA GU	-1.8
AG UC	-1.7	GA CU	-2.3

Free energy changes have been measured for Watson-Crick base pairs (CG, GC, AU, and UA) as well as wobble pairs (GU and UG) [7]. Table 1 shows the change in free energy associated with the addition of a base pair to the helix given an adjacent base pair for Watson-Crick base-pair combinations only. For example, if an AU base pair exists, another A is found to the right of the paired A, and another U is found to the left of the paired U, then 0.9 kcal/mol is released by the formation of two hydrogen bonds between the previously unpaired A and U. This can be seen in the upper left entry of Table 1. Note that swapping the nucleotides in an entry both vertically and horizontally results in a new valid entry. Hence, the bottom right entry $\begin{smallmatrix} G & A \\ C & U \end{smallmatrix}$ with ΔG = -2.3 also implies an entry $\begin{smallmatrix} U & C \\ A & G \end{smallmatrix}$ with ΔG = -2.3.

A single mutation of a base pair to another base pair implies changing two of these propagation energies, one with the mutated pair on the right and another with the mutated pair on the left. Table 2 shows the overall change in free energy for the case where a single base pair mutates from CG to GC (where the first nucleotide specified is nearer to the 5' end of the molecule than the second nucleotide). The un-mutated pair is shown as an italic CG in the table. The four larger bold bases are the consensus neighboring bases (where it is assumed the mutation of interest is happening at a location not on the end of a helix). From the table it is clear that the change in energy from this mutation depends on the neighboring bases in the helix. Furthermore, the differences are not insignificant as can be seen by comparing the magnitudes of the values in Table 2 to those in Table 1. The currently used priors enforce a single substitution penalty for the CG to GC mutation independent of the consensus neighboring bases.

Table 2. Free energy changes for CG to GC mutation[*]

Neighboring Bases	ΔG (kcal/mol)	Neighboring Bases	ΔG (kcal/mol)
5'-A*C*A-3' 3'-U*G*U-5'	-0.1	C*C*C G*G*G	+0.4
A*C*U U*G*A	0.0	C*C*G G*G*C	0.0
U*C*A A*G*U	0.0	G*C*C C*G*G	0.0
U*C*U A*G*A	+0.1	G*C*G C*G*C	-0.4
A*C*C U*G*G	-0.1	C*C*A G*G*U	+0.4
A*C*G U*G*C	-0.5	C*C*U G*G*A	+0.5
U*C*C A*G*G	0.0	G*C*A C*G*U	0.0
U*C*G A*G*C	-0.4	G*C*U C*G*A	0.1

[*] Ignores a further possible change of 0.4 in free energy due to breaking or inducing symmetry in the full helix.

This lack of neighbor dependence can not be overcome without rewriting the parameter estimation code. This rewriting is feasible since the source code is publicly available, but requires significant coding effort. To use the current code, the values in Table 2 could be averaged into a single value (0.0) for CG to GC mutation. Similar values could be obtained for mutations such as AU to GC or UA to GU, etc. (which in general would not be 0.0). If mutations with large positive energy changes are considered less likely (since they are less likely to hold the helix together), then simulated sequences with fewer of these large positive-energy-change events can be generated and used with existing software to estimate priors. If these thermodynamic-measurement based priors have similar effectiveness to those based on generic ncRNA sequences, then perhaps incorporating neighbor dependence will have some advantage.

Table 3 shows an example of the free energy changes and the score changes associated with observed GC to GU mutations in the known Cardiovirus CRE sequences. The mutation positions are the fourth and fifth base pairs from the top in Figure 2. The third and sixth base pairs from the top of the figure are always GC in all 30 known sequences. The first column of the table shows the four observed mutation patterns, with the larger bold symbols in the middle showing the mutating pairs. The second column shows the number of times the mutation pattern is observed. Notice that the top pattern is the consensus even though the second pattern is the most often observed. This is partly due to the fact that the consensus was defined from the twelve seed sequences (Figure 1) used to estimate the model and not from the full set including an additional eighteen sequences found by searching with the model. It is also partly due to entropy weighting, since the overall sequences of the sixteen with the second pattern are much more similar to each other than to other sequences in the full set.

Two sets of score changes are shown, one using Dirichlet mixture priors from version 0.7 of Infernal (ΔScore_{70}) and another using only pseudocounts from version 0.55 (ΔScore_{55}). These changes in scores are relative to the consensus sequence and are found by taking the difference between the GC and GU emission scores. The change in score for the last row is by definition the sum of the score changes in the second and third rows since the two mutations are taken as independent. The third column of Table 3 shows the sum of the three helix propagation energy changes for the helix pattern (one energy for the two left pairs, one energy for the two central pairs, and one energy for the two right pairs). The free energy increases with GC to GU mutations as is to be expected since GU bonds are not as strong as GC. One notices a correlation between free energy changes and score changes. By regressing score changes on free energy changes, appropriate conversion factors between the two might be obtained.

Table 3. Free energy and score changes in Cardiovirus CRE (RF00453) family due to GC/GU mutations

Helix Pattern	Observed	ΔG (kcal/mol)	ΔScore_{70} (bits)	ΔScore_{55} (bits)
G**GC**C C**CG**G	6	-9.2	0	0
G**GU**C C**CG**G	16	-6.1	-2.46	-0.89
G**GC**C C**UG**G	6	-6.1	-2.78	-1.15
G**GU**C C**UG**G	2	-3.1	-5.24	-2.04

3.4 Dependence of Free Energy Changes on Loop Lengths

Thermodynamic measurements of RNA structures indicate that the number of bases in loops (unpaired segments of the RNA sequence) has a significant influence on the stability of the RNA molecule [7]. Three types of common loops will be discussed here: hairpin, bulge, and internal. The hairpin loop is a segment of unpaired sequence directly between two segments that are base paired with each other. A bulge loop is a segment of unpaired sequence that interrupts a helix on one side (either the 5'-side of the helix or the 3'-side). An internal loop is a pair of bulges on both the 3'- and 5'-side of a helix between the same two base pairs of the helix. In Figure 2, a hairpin loop is found at the bottom of the figure and a 5'-bulge loop with three unpaired bases is found on the left side of the figure. The 5'-bulge loop interrupts the helix between two AU base pairs. If there was also a 3'-bulge loop between the two U's, then the pair of bulge loops would instead be classified as a single internal loop.

Table 4. Free energy changes for loop-length changes [7]

Loop Length	Internal	Bulge	Hairpin
1	-	+3.3	-
2	+0.8	+5.2	-
3	+1.3	+6.0	+7.4
4	+1.7	+6.7	+5.9
5	+2.1	+7.4	+4.4
6	+2.5	+8.2	+4.3
7	+2.6	+9.1	+4.1
8	+2.8	+10.0	+4.1
9	+3.1	+10.5	+4.2
10	+3.6	+11.0	+4.3
12	+4.4	+11.8	+4.9
14	+5.1	+12.5	+5.6

All free energy changes in units of kcal/mol

Table 4 shows the free energy changes for various loop lengths for the three types of loops. Internal loops of length 1 are not possible since each of the component bulges must have at least one unpaired base. Hairpin loops of length 1 and 2 result in steric hindrance such that it is impossible to form the hydrogen bonds on the last base pair of the enclosing helix. Even if the last base pair of the potential helix was a Watson-Crick GC pair with very strong three hydrogen bond potential, it will not form, so the resulting length 1 or 2 hairpin loop will actually become a length 3 or 4 loop. Hairpin loops of length 3 or 4 are still very tight and will tend to almost entirely offset the free energy of the end base pairs of the helix. This can be seen by the high +7.4 and +5.9 kcal/mol free energy changes for hairpin loop lengths of 3 and 4. Bulge loops increasingly bend and disrupt the helix as they grow, so the free energy changes increase monotonically with loop length. Internal loops distort the helix much less and therefore have much smaller free energy changes. It should be noted that an internal loop of length two can be generated where no internal loop previously existed by a mutation in a base pair such that the resulting pair is no longer a Watson-Crick or GU pair. Hence the mutation free energy change values discussed in Section 3.3 between Watson-Crick/GU pairs can be augmented to include all possible mutations.

The dependence on loop lengths can be integrated into the CM parameter estimation if a separate prior is allowed for each node of the CM rather than using the same prior for every node in the model of the same type. Rather than have a single prior for all L nodes, separate priors for L nodes in hairpin loops with consensus length 3, 4, 5, etc. need to be specified. The prior for the length-3 hairpin L nodes would have a very large penalty on state transitions through the D state, whereas the prior for length-8 hairpin L nodes would have a much smaller D-state transition penalty.

Table 5. CM scoring for loop-length changes[*]

Loop Change	Context	CM Node	CM State	RF453 ΔScore[**]
Insert Hairpin	L_1-L_2	L_1	IL	-5.29
Insert Hairpin	P-L	P	IL	-5.70
Insert Hairpin	L-P	P or L	IR or IL	-4.70[***]
Delete Hairpin	Any	L	D	-5.08
Insert 3'-Bulge	R_1-R_2	R_2	IR	-
Insert 3'-Bulge	P-R	R	IR	-
Insert 3'-Bulge	R-P	P	IR	-
Delete 3'-Bulge	Any	R	D	-
Insert 5'-Bulge	L_1-L_2	L_1	IL	-5.29
Insert 5'-Bulge	P-L	P	IL	-5.70
Insert 5'-Bulge	L-P	L	IL	-5.29
Delete 5'-Bulge	Any	L	D	-5.08

[*] **Internal loops in a CM are just a 3'-bulge and a 5'-bulge that happen to be internally adjacent to the same P node. There is no explicit linkage or special notation for the pair of bulges.**

[**] **ΔScore is relative to the local consensus sequence. Additional loop-length changes (multiple insertions or deletions) require slightly different transition scores. Score changes are for Infernal 0.55 models.**

[***] **ΔScore using L.IL is -4.70 which exceeds ΔScore of -5.29 of P.IR, so dynamic programming will choose path with -4.70 score contribution.**

Table 5 shows how insertions and deletions are implemented in a CM for the various types of loops and whether the insertion or deletion is internal to the loop or at one of the loop ends. Only hairpin and bulge loops are shown since CMs treat internal loops as two completely independent bulge loops. The context column shows which type of CM node is associated with the consensus sequence symbol to the left and right of the insertion or deletion position. The CM node column shows the node that contains the state responsible for the insertion (IL or IR) or deletion (D) and the CM state column shows the responsible state. In the case of insertions on the 3'-end of a hairpin loop there is ambiguity as to whether this insertion is done by the L node to the left or the enclosing P node (3rd row of table). In practice, whichever of the two has a smaller (in magnitude) score penalty for the insertion is the one that matters. The last column of the table shows the score reductions resulting from the insertion or deletion for the Cardiovirus CRE (RF00453) family when estimated using Infernal 0.55 (without Dirichlet priors).

The insertion and deletion penalties in the last column of Table 5 are substantial. The score of the consensus sequence for this family is 37.75 and the threshold used is 25.0 (called the *gathering cutoff* in Rfam), so the penalties for changes from the consensus sequence can not exceed 12.75 without the database sequence being rejected. The best score for a rejected sequence is 19.44 (*noise cutoff*) and the worst score for an accepted sequence is 27.51 (*trusted cutoff*) for this family in Rfam, so the separation between accepted and rejected database sequences is rather small. Since there are no observed insertions or deletions in any known members of this family, all score penalties in Table 5 depend purely on priors (in this case uninformative priors, since Infernal 0.55 is shown). It is entirely possible that actual family members with insertions and deletions have been missed because the estimated penalties for these insertions or deletions are so high based on the priors. More information on which insertions or deletions are reasonable and which are not might allow priors that could find such cases.

Some of the score changes in the last column of Table 5 are forced to be the same by the current implementation. The four values associated with hairpin loops must be the same as the values for the 5'-bulge (with the exception of the L-P context) since the priors must be the same and no insertions or deletions are observed in either case. There is no mechanism to distinguish between hairpin loops, bulge loops, or internal loops. There is also no way to incorporate information about consensus loop lengths. Modification of the cmbuild program in the Infernal package to allow different priors would allow this. The cmbuild program already has access to the consensus structure, so determination of consensus loop types and lengths by the program would not be very difficult.

4. CONCLUSIONS

The presentation in this paper has been mostly conceptual, with a few illustrations from a real ncRNA family used to demonstrate the points discussed. The intent is to show that there is potential to improve covariance model based ncRNA gene search by including information contained in experimental thermodynamic data into the CM parameter estimation process. This thermodynamic data can be used as an independent source of information contained in current parameter estimation priors and therefore potentially reduce noise through averaging. The data also allows for the production of more specific priors that depend on neighbors in helices and on loop lengths. The current Infernal parameter estimation program (cmbuild) does not allow for such context dependence, but could with some moderate coding effort. No changes would be necessary to the database search program (cmsearch).

5. ACKNOWLEDGMENTS

SFS is grateful for support under NIH Grant Number P20 RR016454 from the INBRE Program of the National Center for Research Resources. KCW would like to thank Dr. Herbert Tsang and Andrew Hendriks for their helpful feedback on an earlier draft of this paper.

6. REFERENCES

[1] Eddy, S., and Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22 (1994), 2079-2088.

[2] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[3] Eddy, S. Hidden Markov models. *Current Op. Structural Bio.*, 6 (1996), 361-365.

[4] Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S., and Bateman, A. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33 (2005), D121-D141.

[5] Rfam. http://rfam.janelia.org/.

[6] Eddy, S. *Infernal 0.81 user's guide*. http://infernal.janelia.org/, 2007.

[7] Freier, S., Kierzek, R., Jaeger, J., Sugimoto, N., Caruthers, M., Neilson, T., and Turner, D. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Nat. Acad. Sci. USA*, 83 (1986) 9373-9377.

[8] Zucker, M. Computer prediction of RNA structure. *Methods Enzymology*, 180 (1989), 262-288.

[9] Wiese, K., Hendriks, A., and Deschênes, A. Analysis of thermodynamic models and performance of RnaPredict - an evolutionary algorithm for RNA folding. *IEEE Symp. Comp. Intell. Bioinformatics Comp. Bio.*, 2006, 343-351.

[10] Nawrocki, E., and Eddy, S. Query-dependent banding (QDB) for faster RNA similarity searches. *PloS Computational Bio., 3*, 3 (2007), 540-554.

[11] Smith, S. Covariance searches for ncRNA gene finding. *IEEE Symp. Comp. Intell. Bioinformatics Comp. Bio.*, 2006, 320-326.

[12] Weinberg, Z., and Ruzzo, W. Faster genome annotation for non-coding RNA families without loss of accuracy. *Int. Conf. Res. Comp. Mol. Bio.*, 2004, 243-251.

[13] Pleij, C. RNA pseudoknots. *Current Op. Structural Bio.*, 4 (1994), 337-344.

[14] Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I., and Haussler, D. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *Comp. Appl. Biosci.*, 12 (1996), 327-345.