7-1-2006

Accelerated Non-Coding RNA Searches with Covariance Model Approximations

Jennifer A. Smith
Boise State University

2006 IEEE Congress on Evolutionary Computation
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Accelerated Non-coding RNA Searches with Covariance Model Approximations

Scott F. Smith, *Senior Member*, *IEEE*
Department of Electrical and Computer Engineering
Boise State University
Boise, ID 83725-2075
sfsmith@boisestate.edu

Abstract- Covariance models (CMs) are a very sensitive tool for finding non-coding RNA (ncRNA) genes in DNA sequence data. However, CMs are extremely slow. One reason why CMs are so slow is that they allow all possible combinations of insertions and deletions relative to the consensus model even though the vast majority of these are never seen in practice. In this paper we examine reduction in the number of states in covariance models. A simplified CM with reduced states which can be scored much faster is introduced. A comparison of the results of a full CM versus a reduced-state model found using a genetic algorithm is given for the *let7* ncRNA family.

1 Introduction

Covariance models are a very powerful but extremely slow method for searching genome databases for potential non-coding RNA genes. The models are statistically based, with probabilities of insertions, deletions, and mutations estimated from the observed frequencies of events in a known family of ncRNA genes. The slowness of the algorithm can be traced to the attempt to fit sequences of all lengths between 1 and a model-specific maximum length D to every possible contiguous sub-model of the family. Any consensus nucleotide position may be deleted and insertions of any length (such that the overall scored sequence is at most length D) can occur at any position.

The recent discovery of a lossless prefilter to eliminate significant portions of the database sequence from consideration by the CM [1] has not reduced the need to improve the speed of covariance models. If anything the need for faster covariance models has increased since the prefilter has changed the necessary computation time from unaceptable to merely much too long.

The importance of finding non-coding RNA genes has increased as molecular biologists have found increasing numbers of catalytic and regulatory mechanisms which are performed by functional RNA molecules alone or ncRNA molecules in conjunction with proteins [2]. These ncRNA genes include the transfer RNAs (tRNA) that transport amino acids, ribosomal RNA (rRNA) which performs an essential role in protein synthesis, and a component of telomerase which is associated with maintenence of chomosome ends [3].

In this paper we attempt to find simplified covariance models which score well against the members of a ncRNA family. The simplified models have a reduced number of strategically placed insert and delete states. The attempt is to cover the few major variations in structure that might occur in a family and let the minor variations result in small degradations in search efficacy. In this way, small losses in specificity and/or sensitivity are traded for gains in search speed.

The method used to find the simplified models is to use a genetic algorithm since the choices of insert and delete states to retain interact with each other and thus there is no simple way of choosing whether to retain a given state from the original model independent of the others.

Section 2 provides an overview of how covariance models are used to search for ncRNA genes. The simplified covariance models are described in section 3. The method for finding a good simplified covariance model using a genetic algorithm is described in section 4. Experimental results are shown in section 5 for a specific ncRNA family (the *let7* ncRNA family). Section 6 concludes.

2 Covariance Models

A covariance model [4] is a statistical model of nucleotide sequences belonging to a family of non-coding RNAs. The model is based on a multiple alignment of sequences where the alignment columns are annotated with base pairing information. Like a profile hidden Markov model (HMM), observed frequencies of symbols and gaps in alignment columns are used to estimate probabilities of symbols at each location as well as probabilities of insertions and deletions at each location. Unlike an HMM, a covariance model also specifies the probabilities of long-range interactions within the single-stranded RNA three dimensional structure. This is possible because the

0-7803-9487-9/06/$20.00/©2006 IEEE

covariance model is a context-free grammar, whereas the HMM is based on a regular grammar incapable of modeling long range interactions [5].

A CM has six different types of nodes: P, L, R, B, S, and E. A pair-wise emission node (P) specifies that the consensus structure of the ncRNA family contains a base pair between two columns of the multiple alignment. Information associated with this type of node include the sixteen probabilities of emitting each of the sixteen possible pairs of nucleotides, the probability that the base pair is omitted, the probability that only the 3' nucleotide of the base pair exists, the probability that only the 5' nucleotide exists, and the probabilities that additional nucleotides are inserted between the consensus base pair. A left (L) or right (R) emission node specifies an unpaired base in the consensus structure. Probabilities associated with L and R nodes include the probabilities of emitting each of the four possible nucleotides, the probability that the position is omitted, or that one or more additional inserted nucleotides are emitted next to the consensus position. A bifurcation node (B) allows the joining of two sub-models which are contiguous along the sequence. A start (S) node is used at the head of any branch (including the root) and an end node (E) is used to end a branch at a leaf position.

Figure 1 shows an example of a multiple alignment annotated with structural information. The "-" symbol indicates a non-based-paired alignment column. The ">" and "<" symbols mean that the column is base paired with a column to the right and left respectively. This notation only works for structures without pseudoknots. Since covariance models can not handle pseudoknotted structures, this limitation of the notation is not a problem. When a structure actually has a pseudoknot, some of the base pairing information is ignored by the covariance model and the associated bases treated as if they were unpaired. This results in some loss of power in database search. In figure 1, the second column is base paired with column ten. The "." symbol means that the column does not exist in the consensus structure. The example data used in Figures 1, 2, 3, and 5 are purely fictitious and are provided only to demonstrate the mechanics of covariance models.

```
Rat                 AUGG.ACCAAG.GUCAGAC
Bat                 AUGAACUCCAGCGUCCGAC
Cat                 CGG..GUCCCG.GA.AUU.
Fly                 A.GAACUCG.G.GUCAGAC
Cow                 AUCA.UUGUAG..UUA.AC
Consensus:
Structure           ->>>.-<<-<-.>>---<<
Sequence            AUGA.CUCCAG.GUCAGAC
```

Figure 1: Multiple alignment annotated with structure

Figure 2 shows the consensus sequence and structure of the ncRNA family. The dots indicate that two positions are base paired. The "A" pointed to by the arrow labeled "L1" is the first nucleotide in the consensus sequence (*i.e.* it is on the 5' end of the molecule). It will be assigned to an L node with node index 1. The dotted pair "AU" with the dashed oval around it will be assigned to a P node with index 7. A bifurcation is necessary to split the two groups of pair nodes (the two stems).

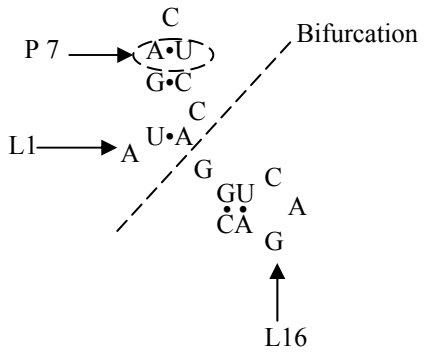

Figure 2: Consensus sequence and structure

Figure 3 shows the nodes of a covariance model based on the multiple alignment in figure 1. The CM forms a binary tree with an S node at the root (called the root start node). The consensus sequence characters are shown next to the CM node that emits them (P, L or R). We note that some of the columns of the multiple alignment could be represented with either an L or an R node. When this happens, an L node is always used by convention.

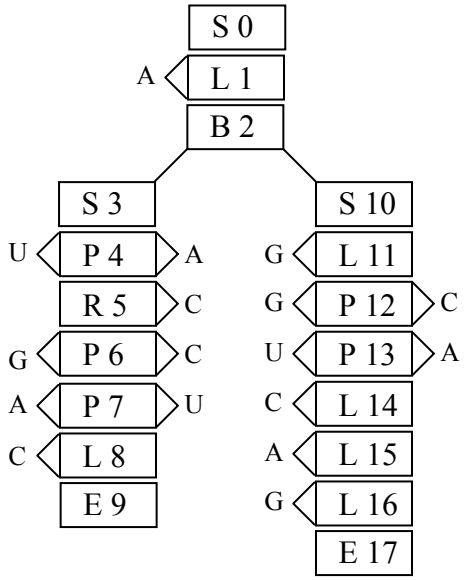

Figure 3: CM nodes in binary tree

The nodes of the model contain one or more states. For the emitting L, R, and P nodes, one of the states represents a consensus character (L and R nodes) or characters (P node) associated with the node and the other states allow for insertions and deletions relative to the consensus. Figure 4 shows the detail of an L node. These nodes contain a match left (ML) state which represents the consensus structure emission of an unpaired nucleotide. In the figure, the L8 node is shown in particular, so the ML8 state emits the consensus character C with highest probability. The other nucleotides (A, G, and U) are emitted with a lower, but nonzero, probability.

Deletions and insertions relative to the consensus are accomplished by either bypassing ML8 via the non-emitting D8 state or passing through the emitting IL8 state in addition to the ML8 state. The IL8 state has a self loop which allows for an arbitrary number of insertions to the right of the match character.

The internal state structure of all nodes has either one or two tiers. The top tier always exists and contains all states other than insert states and the bottom tier contains all insert states. All states of a node are connected to the top tier states of the node below. In Figure 4, the thick arrows indicate (possibly) multiple connections. The ML8 and D8 states are in the top tier and the IL8 state is in the bottom tier. It is the potential elimination of the D8 and IL8 states within the L8 node that is investigated in this work. All nodes of the original CM will be retained and all nodes will retain the state associated with the consensus nucleotide or nucleotides.

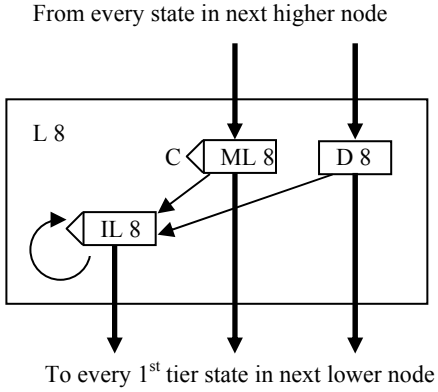

Figure 4: Internal state detail of a left node

A search of a chromosome nucleotide sequence of length L for possible ncRNA genes belonging to the ncRNA family associated with a CM can be done using the CYK algorithm [4]. The algorithm starts at the end states at each leaf of the binary tree describing the CM and works up the tree toward the root. Linear programming is used to find the best possible score for every subsequence ending at each position j in the chromosome sequence and of each possible subsequence length d up a to maximum subsequence length of D. The states are given a unique index number v such that $\gamma(v, j, d)$ represents the best possible score for the chromosome subsequence of length d ending at j fitted to the sub-model including state v and all states below it in the tree. The start state in the root start node is given index 0, so $\gamma(0, j, d)$ is the overall score of the model and the maximum of this score over d represents the likelihood that a gene of the family is located in the chromosome ending at sequence position j.

End (E) nodes contain only one state called an end (E) state. These states take a score of 0 for the null sequence ($d = 0$) at all positions along the chromosome. The E states are not allowed to hold any symbols, so the score for $d > 0$ is minus infinity:

$$\text{E: } \gamma(v, j, 0) = 0;\ \gamma(v, j, d) = -\infty, \text{ where } d > 0 \quad (1)$$

Pair-wise emission (MP) states appear only as the consensus emitting state in P nodes. Since the state emits two symbols (one on each end of the sequence passed up from below in the tree), the resulting sequence with two added symbols can be no shorter than length two. Therefore, the score is set to minus infinity for $d = 0$ and $d = 1$. Otherwise, the log likelihood e of emitting the two characters found in the chromosome sequence is added to the best child state score. The symbol emitted on the right is located at position j and the symbol emitted on the left is located at position $j+1-d$. These symbols are denoted $x(j)$ and $x(j+1-d)$ respectively. The score from a child state y is the child state's score $\gamma(y, j-1, d-2)$ plus the transition probability from the child to the MP state $t(v, y)$. The child state's score is taken for a subsequence of length two shorter and starting one position earlier in the chromosome sequence to allow room for the two symbols emitted by the pair-emitting state:

$$\text{P: } \gamma(v, j, 0) = \gamma(v, j, 1) = -\infty; \quad (2)$$
$$\gamma(v, j, d) = \max_y [\gamma(y, j-1, d-2) + t(v, y)] + e[v, x(j+1-d), x(j)],$$
$$\text{where } d > 1, \text{ and } y \text{ is a child of } v$$

Left emission (ML, IL) states appear as the consensus emitting states of L nodes (an ML state), as states to add inserted symbols relative to the model in P, L, and some S nodes (an IL state) and as a partial match in P nodes where the right portion of the emitted pair is deleted (an ML state). Since this state emits one symbol, the resulting sequence out of the state can be no shorter than length one. Therefore the score of the null sequence ($d = 0$) is set to minus infinity. Otherwise, the log likelihood of emitting the single symbol is added to the best score from the child states:

$$\text{L: } \gamma(v, j, 0) = -\infty; \quad (3)$$
$$\gamma(v, j, d) = \max_y [\gamma(y, j, d-1) + t(v, y)] + e[v, x(j+1-d)],$$
$$\text{where } d > 0, \text{ and } y \text{ is a child of } v$$

Right emission (MR, IR) states appear as consensus emitting states of R nodes, as insertion states in P, R, and the root S node, and as a match state in P nodes where the left symbol of the pair has been deleted. This state emits a single symbol and therefore the score of the null string is minus infinity. The scoring of non-null sequences is similar to that for the left emission states:

$$\text{R: } \gamma(v, j, 0) = -\infty; \qquad (4)$$
$$\gamma(v, j, d) = \max_y [\gamma(y, j-1, d-1) + t(v, y)] + e[v, x(j)],$$
where $d > 0$, and y is a child of v

Bifurcation (B) states consider all possible divisions of d symbols from the chromosome between two branches of the CM binary tree. The left-most $d-k$ symbols are assigned to the left branch and k symbols are assigned to the right branch. The value of k which generates the best overall score is chosen:

$$\text{B: } \gamma(v, j, d) = \max_{0 \le k \le d} [\gamma(y, j-k, d-k) + \gamma(z, j, k)], \qquad (5)$$
where y and z are the children of v

Delete and start (D, S) states are place holders that allow different transition probabilities to be attached for different child states. Delete states appear in P, L, and R nodes and serve as a path around the match state. S states appear in S nodes and serve to collect multiple paths together into a single path either at the root or as one of the two branches of a B node:

$$\text{D or S: } \gamma(v, j, d) = \max_y [\gamma(y, j, d) + t(v, y)], \qquad (6)$$
where y is a child of v

Since all states in the CM are evaluated for every position j and length d in the chromosome database sequence, eliminating model states will significantly reduce computational time. The elimination of insert and delete states is considered in what follows. This includes consideration of all IL, IR, and D states. It also includes consideration of ML and MR states within P nodes.

3 Simplified Covariance Models

When looking at the structure-annotated multiple alignments of ncRNA families one sees that there are many columns that contain no dots (no deletions relative to the consensus model) and rather few columns with a dot as the consensus character (few insert locations relative to the consensus model). The standard covariance model allows any consensus column to be deleted and any number of inserts between any two adjacent columns. The number of states in the covariance model could be greatly reduced if the delete and insert states associated with strongly conserved portions of the sequence were eliminated.

```
Rat              AUGG ACCAAG.GUCAGAC
Bat             AUGAA CUCCAGCGUCCGAC
Cat              CGG. GUCCCG.GA.AUU.
Fly              AGAA CUCG.G.GUCAGAC
Cow              AUCA UUGUAG..UUA.AC
Consensus:
Structure        ->>> -<<-<-.>>---<<
Sequence         AUGA CUCCAG.GUCAGAC
```

Figure 5: Removal of the insert state to right of "AUGA"

The extent to which removal of a delete or insert state is detrimental to the score of sequences which are true family members depends in part on how many sequence characters are forced out of their optimal alignment positions. It is possible that a sequence which needs an insert which has been removed from the model also uses a delete in close proximity. In this case, if both the insert and delete are not used, then the number of shifted characters may be small. An example of this is given in Figure 5 where we have removed the first insert state from Figure 1. The fly sequence has only lost the "GA" portion of the original multiple alignment. The subsequence "GAA" has moved to the left, but the final "A" was not initially matched (it was in the column with consensus insert initially - a "." consensus structure).

In the case of bat in Figure 5, there is no nearby delete in the sequence to absorb the character from the missing insert state, so the entire subsequence "AUGAA" gets moved to the left and the matching of "AUGA" is lost. It is difficult to come up with a simple rule for when to allow the removal of an insert or delete state based solely on the number of sequences using a particular state in the optimal multiple alignment. To find a reasonable set of states to retain, we will use a genetic algorithm to search the space of retained-state combinations with the dual objectives of keeping the number of retained states high and the reduction in mean score of the known ncRNA family low.

4 Finding Retained States of Simplified CMs

The determination of which insert and delete states of the original covariance model to retain can be made using a genetic algorithm (GA). A binary vector where 1 means retain a state and 0 means omit a state is used. For each P, L, or R node of the original model, four, two, or two bits respectively appear in the binary vector. For a P node, the bits represent IL and IR states and DR and DL pseudostates. If both DR and DL is omitted, then the D, MR, and ML states are removed. If only DL is omitted, then only MR is removed. If only DR is omitted, then only ML is removed. For an L node, the bits represent D and IL states. For an R

node, they represent D and IR states. Using a representation with delete and insert retention choices co-mingled allows good local choices of delete/insert retention to be retained during cross-over. Table 1 gives an overview of the entire representation.

TABLE 1
SUMMARY OF GA REPRESENTATION BITS BY NODE TYPE

Node Type	Insert States	Delete States	GA Bits
P	IL, IR	D, MR, ML	IL, IR, DR*, DL*
L	IL	D	IL, D
R	IR	D	IR, D
S (root)	IL, IR	-	IL, IR
S (right bifurcation)	IL	-	IL

*DR and DL are pseudostates. If DR and DL are 0, then D, MR, and ML are removed. If only DR is 0, then only ML is removed. If only DL is 0, then only MR is removed.

The vector with all 1s should always be included in the initial generation of the GA. This is simply a representation of the original model. Placing the original model in the initial generation makes sure that the possibility of no state reduction is considered.

The fitness function must penalize inclusion of more states while encouraging higher mean scores for known ncRNA family members. The fitness function $f = s - \alpha n$ can be used where f is the fitness, s is the mean score, n is the number of retained states, and α is a positive adjustment parameter. The parameter should be chosen larger for more aggressive state reduction and smaller for less loss of score. The choice of $\alpha = S / 2N$, where S is the mean score and N is the number of states in the original model is reasonable if the maximum loss of score that can be tolerated is half of the original score. In this case $f = S/2$ for the original model. The lowest possible score for any combination of retained states which improves on the original model is $S/2$ (which occurs when $n = 0$, and $s = S/2$ plus a small amount). As long as the best individual always is passed to the next generation such that monotonically non-decreasing best fitness is assured, then $S/2$ is the smallest possible score.

5 Experimental Results

To test the idea of simplifying a covariance model by reducing the allowed states in the model nodes, the *let7* ncRNA [6] covariance model was investigated. The original model was taken from the rfam ncRNA database [7]. The family of this model has 47 known members and is composed of 31 pair nodes, 11 left nodes, and 10 right nodes. The structure of the model is shown in Figure 6. The notation "*n*P>" means that the consensus sequence has a group of *n* columns assigned to P nodes where the paired columns are to the right in the sequence. The notation "*n*L" is for a group of *n* L nodes and "*n*R" for a group of *n* R nodes. Using the same notation, the structure of Figure 1 could have been described as "1L 3> 1L 2< 1R 1< 1L 2> 3L 2<".

```
6P> 1L 21P> 3L 4P> 7L 4P< 10R 27P<
```

Figure 6: Structure of *let7* covariance model

The scores that are obtained by fitting the 47 known *let7* ncRNA sequences to the match states of the covariance model are shown in Table 2. The sequences are optimally aligned to the match states since all nodes have a full set of insert and delete states. The "pairs" scores are those obtained from the match states of the 31 pair nodes only. The "singles" scores are those from 21 left and right nodes only. The "pairs and singles" are from all 52 nodes. The units of the scores are bits and represent the $\log_2$ of the likelihood ratio for the fitted sequence compared to a random sequence. Therefore a threshold of 40 bits on the score during a search would find all 47 sequences with an expected false positive rate of one per 1.1×10^{12}. The minimum and maximum columns do not necessarily add as the mean column does since the minimums and maximums do not necessarily occur in the same sequence.

TABLE 2
SCORES WITH COMPLETE INSERT/DELETE COVERAGE

	Mean	Minimum	Maximum
Pairs	58.16	38.08	65.80
Singles	13.03	5.92	20.52
Pairs and Singles	71.19	49.71	82.97

Table 3 shows scores comparable with those of Table 2 for the case where inserts and deletes have been limited to three inserts and 17 nodes with deletes. In the original model, all 52 nodes can be deleted and all possible inserts can occur. A threshold of 40 bits using this reduced set of insert and delete states would still find all 47 sequences.

TABLE 3
SCORES WITH LIMITED INSERT/DELETE COVERAGE

	Mean	Minimum	Maximum
Pairs	56.74	28.81	65.80
Singles	12.49	4.24	20.52
Pairs and Singles	69.23	44.49	82.53

The details of the delete states and insert states retained are given in Table 4 and Table 5 respectively. The nodes that are chosen to include delete states are mostly single emission nodes (L nodes or R nodes), with only one pair node having a delete state. Table 2 shows that the last node of the "21P>" and the first node of the "27P<" group have a delete state, but these are the two halves of a single pair

node. The result is that 35 out of 52 delete states have been eliminated. Of the 62 total MR and ML states in the P nodes only two remain. Only three insert states have been retained and they are all on the ends of single emission groups. The "7L" group retains a single insert on its left and on its right ends. The "10R" group retains a single insert on its left end. Since the bulk of the model's power comes from the pair nodes and gaps rarely appear in the multiple alignments within pair node groups, the main function of the insert states seem to be to keep the adjacent pair node groups properly aligned. Additional investigation of whether the left and right nodes can be eliminated altogether might result in even simpler yet still effective models.

TABLE 4
NODES CHOSEN WITH DELETE STATES FOR LIMITED COVERAGE CASE

Structure Group	Nodes	Structure Group	Nodes
6P>	None	7L	First four
1L	None	4P<	None
21P>	Last one	10R	All but 4th
3L	All	27P<	First one
4P>	None		

TABLE 5
NODES CHOSEN WITH INSERT STATES FOR LIMITED COVERAGE CASE

Structure Group	Nodes	Structure Group	Nodes
6P>	None	7L	Left of first
1L	None		Right of last
21P>	None	4P<	None
3L	None	10R	Left of first
4P>	None	27P<	None

6 Conclusions

It appears that traditional covariance models that include insert and delete states in every emission node may be significantly over-parameterized. We have seen that eliminating most of these states from the *let7* ncRNA covariance model causes very little detriment to the effectiveness of the model.

Further investigation should include experimenting with more ncRNA family models to see if a large amount of state reduction is generally possible. A quick look at the multiple alignment of other families leads one to believe that this is likely the case since there are often large gap-less blocks, especially in blocks assigned to pair nodes. It is also worth investigating whether models including only pair nodes are sufficiently powerful for ncRNA gene database search.

Acknowledgments

The project described was supported by NIH Grant Number P20 RR016454 from the INBRE Program of the National Center for Research Resources.

Bibliography

[1] Z. Weinberg and W. Ruzzo, "Faster Genome Annotation of Non-coding RNA Families Without Loss of Accuracy," *Int. Conf. on Research in Computational Molecular Biology*, pp. 243-251, 2004.

[2] R. Gesteland, T. Cech, and J. Atkins, *The RNA World*, 3rd Ed., Cold Spring Harbor Laborartory Press, 2005.

[3] T. De Lange, V. Lundblad, and E. Blackburn, *Telomeres*, 2nd Ed., Cold Spring Harbor Laborartory Press, 2005.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press, 1998.

[5] N. Chomsky, "On Certain Formal Properties of Grammars," *Information and Control*, Vol. 2, pp. 137-167, 1959.

[6] A. Rougvie, "Control of Developmental Timing in Animals," *Nature Review Genetics*, Vol. 2, No. 9, pp. 690-701, 2001.

[7] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. Eddy, "Rfam: An RNA Family Database," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 439-441, 2003.