5-2-2010

Computation Intelligence Method to Find Generic Non-Coding RNA Search Models

Jennifer A. Smith
Boise State University

 DOI: 10.1109/CIBCB.2010.5510341

Computation Intelligence Method to Find Generic Non-coding RNA Search Models

Jennifer A. Smith

***Abstract*—Fairly effective methods exist for finding new non-coding RNA genes using search models based on known families of ncRNA genes (for example covariance models). However, these models only find new members of the existing families and are not useful in finding potential members of novel ncRNA families. Other problems with family-specific search include large processing requirements, ambiguity in defining which sequences form a family and lack of sufficient numbers of known sequences to properly estimate model parameters. An ncRNA search model is proposed which includes a collection of non-overlapping RNA hairpin structure covariance models. The hairpin models are chosen from a hairpin-model list compiled from many families in the Rfam non-coding RNA families database. The specific hairpin models included and the overall score threshold for the search model is determined through the use of a genetic algorithm.**

I. Introduction

COVARIANCE models (CMs) [1, 2] have been quite effective in finding new non-coding RNA (ncRNA) genes in genomic databases. An example of this success is the Rfam ncRNA database [3-5], which uses the Infernal [6, 7] implementation of covariance models for database search. However, the method used by Infernal is only capable of finding new members of known ncRNA families. The method requires a secondary-structure annotated multiple alignment of known family members in order to determine model structure and parameter values.

Often there are an insufficient number of known sequences to reasonably estimate covariance model parameters. As a result, there is heavy reliance on the use of priors in CM parameter estimation. These priors are estimated from a large collection of known ncRNA sequences that are not specific to the family being modeled [8, 9]. In this sense, there is already some generic knowledge of how ncRNA structures evolve embedded in the search models.

A significant drawback of covariance models is the very large computational burden that CM-based search requires. Since there are currently no good generic ncRNA gene search programs available, the entire genome of a newly sequenced organism needs to be searched with every family model (1371 models as of December 2008 and growing rapidly) in order to fully annotate it with the ncRNA genes. In order to make this approach feasible, the genome needs to be pre-filtered to reduce the database portions searched by several orders of magnitude. This is done with primary structure based homology search algorithms such as BLAST [10, 11], which is fast but rejects true positives, or a lossless HMM-based [12] method that is slower and often does not reduce the database enough [13]. A relatively efficient generic ncRNA gene search algorithm would alleviate this.

Determining which sequences belong in a family is problematic. Grouping several sub-families together results in more information, allowing better parameter estimates. The larger groups also allow modeling of the diversity observed between the sub-families and may allow new family members with different combinations of the diverse features than that observed in the original training set to be found. On the other hand, essential features of a particular sub-family may be diluted by mixing with other sub-families that do not require that feature. Attempting to form models of families and component sub-families expands the total number of family models and contributes to the already very large computational burden.

Generic gene finding algorithms for protein-coding genes rely on a variety of signals such as the presence of promoter sequences, open reading frames of appropriate sizes and composition biases within the gene relative to the overall genome [14]. None of these signals are applicable to ncRNA genes. The primary aspect of ncRNA families used to find new genes is the presence of similar secondary structure patterns, with secondary reliance on primary structure (since sequence homology is much weaker for ncRNA genes than for protein-coding genes). A generic ncRNA gene search algorithm mostly likely will have to rely on finding secondary structure components that are similar to generic secondary structure components of known ncRNA genes.

The common feature of most known ncRNAs is the existence of hairpins composed of a intramolecularly base paired stem and an unpaired loop [15]. It is possible to quantify the most common loop lengths and stem lengths. A search could then be undertaken to find locations in the genome capable of forming generic hairpin structures with reasonable stem and loop lengths. Unfortunately, it has already been shown that the number of such structures which are expected to occur at random is far too large for any reasonable false alarm rate [16].

Manuscript received December 22, 2009. This work was supported in part by the U.S. National Institutes of Health under Grant R15GM087646 and Grant P20RR016454.

J. A. Smith is with Boise State University, Dept. of Electrical and Computer Engineering MS-2075, 1910 University Ave., Boise, ID 83725-2075 USA (phone: +1-208-426-5743; fax: +1-208-426-2470; e-mail: jasmith@boisestate.edu).

978-1-4244-6766-2/10/ $26.00 © IEEE

The tactic taken here is to try to split the difference between family-specific models and completely generic models. We will rely on finding groups of hairpin structures that are reshuffled pieces of known ncRNA families. As mentioned earlier, the use of priors already makes the models of some of these structures somewhat generic. We will reject those groupings of known hairpin structures that are too generic such that false alarm rates are too high. We will also reject those groupings that are too specific such that only members of families that contributed these structures are found. The hope is to get a more generic model that finds members of many of the known families as well as members of yet-unknown families in a single pass of the genome database. Further, this model should require no more computational search effort than using covariance model search for a single family.

In the following section, we will describe the generic search model in more detail. This is followed by a description of the genetic algorithm used to select hairpin components and thresholds. Finally, some experimental results showing the potential of this method are presented before a few concluding remarks.

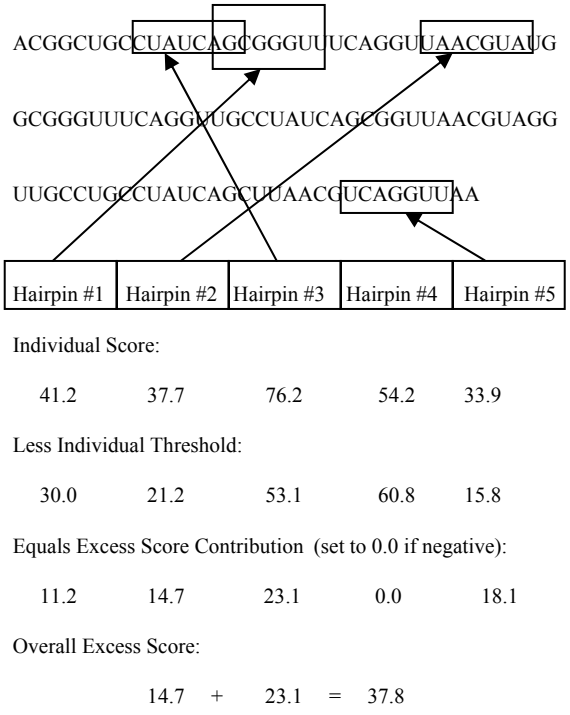

Fig. 1. Scoring a database segment with respect to a search model with five hairpins.

II. Generic ncRNA Search Model

Rather than search for purely generic hairpin loops structures, which Rivas and Eddy [16] have already shown has insufficient specificity, we will search for combinations of known hairpin structures in hopes that some yet unknown ncRNA families have evolved from known families and therefore have pieces that have some small amount of primary and secondary structure homology to known ncRNA hairpin structures. Some of the known structures are already pretty close to generic due to the fact that few family members are known and parameter estimates relied heavily on generic priors. Other known structures are very specific due to large numbers of known family sequences with sufficient variation to get good parameter estimates with little reliance on generic priors. The genetic algorithm based model building of the next section will automatically choose models with a good balance between generic and specific structures.

The search model is composed of a collection of hairpin structures, where hairpin sequence order in the searched database is not important. Each hairpin is a covariance model taken directly from a portion of an Rfam database family model. Each hairpin model has its own individual threshold. A segment of database is scored by scoring each of the search model hairpin models against the segment separately and then determining which combination of hairpin models are most representative of the segment. If the best placement of two search model hairpins within the database segment overlap or are too close together, then only the hairpin with highest excess score is used (too close is a user defined parameter). The excess score is defined as the difference between the individual hairpin score and its individual threshold. The excess score of an individual is set to zero if the value is negative. If the highest-scoring placement of two hairpin structures is too far apart (another user-defined parameter), then the hairpins are divided into groups such that each group is compact enough to not violate the maximum distance parameter. The excess score of the highest excess-scoring group becomes the overall segment excess score.

Figure 1 shows an example search model and its scoring against a segment of database. The example search model has five hairpin structures. The covariance models of each of these structures are individually scored against this database segment (perhaps using the Infernal package ***cmsearch*** program) and the best-scoring location within the segment of each hairpin structure recorded. In the figure, arrows show which database location best maps to each hairpin structure. Hairpin structure #4 has a score of 54.2, but this structure has a threshold of 60.8, so hairpin #4 is deemed not to be present in this database segment.

The best mappings for hairpin #1 and #3 overlap in the database sequence. Since hairpin #3 has a higher excess score, hairpin #3 is deemed to be potentially present rather than hairpin #1. Hairpin #5 is too far away from the other hairpins (as determined by a user-selectable parameter), so we form multiple groups such the largest groups possible can be made without violating this distance constraint. A

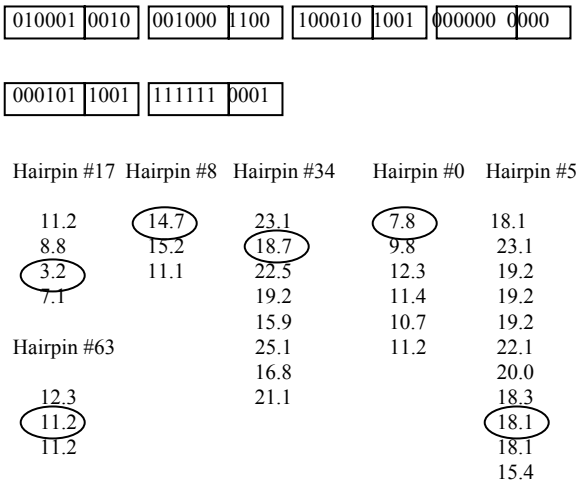

Fig. 2 Representation of search models for genetic algorithm.

given hairpin may be a member of more than one group, but this is not shown in the example. The group composed of hairpin #2 and #3 has the highest sum of excess scores (14.7 + 23.1 = 37.8, versus the group with hairpin #5 only = 18.1), so 37.8 is the overall excess score. The overall excess score is then compared to a user-selectable overall excess score threshold to determine if this database segment is an ncRNA gene hit.

The idea here is to select a subset of the search model's hairpin structures as potentially present and to choose scores and thresholds based only on this subset. However, if the selected subset is not powerful enough, the sum of scores will not exceed the fixed threshold. This will clearly make it difficult to find novel families of very short ncRNA genes, but not doing it would result in unacceptable false alarm rates.

The covariance models for the individual hairpins are quite small. It can be shown that the computational complexity of using a covariance model for database search is between $O(n^2)$ and $O(n^3)$ [17], where n is the length of the consensus sequence of the structure being modeled. The consensus sequences of the individual hairpins are very short (usually between about 10 and 25, compared to many Rfam family models that are in excess of 100), so even if the search model has five to ten of these individual hairpins, the computation time is comparable to running a single average-sized Rfam family covariance model.

III. GA Method to Build Search Model

A list of hairpin covariance models is taken from the Rfam database (for the results below - version 9.1, released December 2008, containing 1371 families). Each model has a list of scores from the training sequences used to construct the model. The easiest way to obtain the hairpin models and scores is to select only those columns of the secondary-structure annotated multiple alignment of training sequences that correspond to the hairpin of interest. Attempting to take apart the covariance model parameter files and renumber nodes and states is much more difficult. The ***cmbuild*** program in the Infernal package is used to create the covariance model parameter files and the ***cmsearch*** program of the same package used to find the scores of the training sequences. Infernal version 1.0.2 released October 2009 was used in the experimental results below.

An individual in the GA is composed of one or more fixed length genes. Complete genes are always created or destroyed by insertion or deletion events. Crossover always takes place at gene boundaries. Each gene is composed of two halves, the first half is a hairpin model index and the second is a training sequence score index. The model index selects a covariance model for inclusion in the set of hairpins in the search model and the score index is used to select a training sequence score which then becomes the individual hairpin threshold as described in the previous section. Figure 2 shows and example individual.

In the example, there are six genes with ten bits each. The first six bits of each gene specify one of 64 possible hairpin structures (indexed 0 to 63) and the last four bits specify one of sixteen possible individual thresholds (indexed 0 to 15). If there are less than 64 hairpins in the hairpin list or less than sixteen scores associated with a particular hairpin, the index value is used in modulo the number of hairpins or scores, so it is never possible to have an invalid GA individual. If the hairpin list has 64 hairpins, then the GA individual shown represents a search model containing hairpins 17, 8, 34, 0, 5, and 63 from the hairpin list, with individual thresholds of 3.2, 14.7, 18.7, 7.8, 18.1, and 11.2 respectively. Since these six hairpins have score lists of length 4, 3, 8, 6, 11, and 3 respectively, some of the score indices had to recycle to the top of the score list (sometimes more than once).

The fitness of an individual is calculated using the number of known ncRNA genes in the training set found (true positives) and the number of hits on one hundred reshuffled versions of the training set (false positives). The fitness function is then $F = t - f * (s/2)$, where t is the number of true positives, f is the number of false positives, and s is the number of sequences in the training set. s is also the maximum possible number of true positives, so any individual with no false positives is always better than any individual with two false positives. In the experimental results that follow $s = 595$.

IV. Experimental Results

A training set was built for 64 randomly selected families out of the 1371 families in version 9.1 of Rfam. If a family model did not include any hairpins (which is rare), then the

TABLE I
PROPERTIES OF FAMILIES INCLUDED IN SEARCH MODEL

Accession Number	Number of Seed Sequences	Consensus Number of Stem Pairs	Consensus Loop Length
RF00115	8	5	12
RF00340	6	5	6
RF01316	20 (16 used)	4	3
RF00552	18 (16 used)	6	6
RF00215	28 (16 used)	8	8
RF01225	26 (16 used)	4	8

TABLE II
CONSENSUS SEQUENCES OF FAMILIES INCLUDED IN SEARCH MODEL

Accession Number	Consensus Sequence (Stems Underlined)
RF00115	GUUGCCAAUUUUCUUCAGUGAC
RF00340	CAGGGCAGCCUCCCUG
RF01316	ACAACUCUUGU
RF00552	AGCUGCAGCGAAGCAGCU
RF00215	ACAGACUCUUCCAGUCUGAGUUUGU
RF01225	GGCCUUGACCGUGGCC

selection was discarded and another family selected. When a family model contains more than one hairpin, exactly one hairpin is selected at random within the family. The result is a list of 64 hairpin structures each with a covariance model and a list of training sequence scores. If there were fewer than 16 family members used to build the covariance model (the 'seed' sequences in Rfam terminology), then scores were listed for every seed sequence. If there were more than 16 seed sequences for a family, 16 sequences are selected at random. Therefore a four-bit score index is sufficient in each individual. If a family has fewer than 16 seed sequences, the list is simply reused from the top when the score index is too big. So, each gene in the GA has a length of ten bits, with six bits for the hairpin index and four bits for the score index.

Three different test sets were examined. The first is a very incestuous test set composed of the sequences which generated the score lists for each of the 6 hairpins in the search model. The second set contained all of the sequences from all 64 hairpins in the hairpin list. The first two sets contained only those portions of the sequences associated with the hairpin structures in the hairpin structure list. The third test set used the full-length sequences of the sequences in the second test set. The fourth test set is a true test set in the sense that 64 families from the 1307 families not selected to build the search model are chosen at random (discarding any that do not contain any hairpin). Up to 16 sequences within each family was chosen at random in the same manner as for the training set. The resulting test set is similar to the third test set in number of sequences and average length of sequences, but is guaranteed not to have any sequence information used to form the search model.

TABLE III
TRUE POSITIVE AND FALSE NEGATIVE COUNTS IN TEST SETS

Test Set	Number of Sequences Over Threshold	Number of Sequences in Test Set	Number of Hits on Randomized Sequences
1	68	78	0
2	294	595	0
3	381	595	1
4	92	561	0

Test set 1 is composed of the 78 sequences in the 6 families that form the search model (8 from RF00115, 6 from RF00340, and 16 randomly selected from each of RF 01316, RF00552, RF00215, and RF1225). Test set 2 is composed of the 595 sequences in all 64 training families (the 6 that ended up in the search model plus the 58 that did not). Both sets 1 and 2 use only the multiple alignment columns associated with the single hairpin selected for the hairpin structure list. Test set 3 contains the same sequences as test set 2, but using the full-length sequences. Test set 4 is generated in the same manner as test set 3, but with families and sequences chosen randomly from those families not included in the training set. Test set 4 is the only true test set in the usual meaning of the term.

A search model was found using the GA method described in the previous section using a population size of 100 and 100 generations. Six hairpins were selected by the GA to form the search model. These are shown in Table I. Stems of lengths between four and eight pairs were selected and loops of lengths between three and twelve. Table II shows the consensus sequences for the six hairpins.

The results of using the search model found on the four test sets described above is shown in Table III. There were 78 sequences in the first test set since the six hairpins selected were associated with 8, 6, 16, 16, 16, and 16 seed sequences resulting in the same number of scores in the score lists. The 68 hits can easily be explained as the GA tending to select scores on the low end of the lists to be used as thresholds. Since the sequences were so short, there was never more than one non-overlapping hairpin hit and the best score was always for the hairpin in the search model that matched the family of the test set sequence.

In test set 2, 294 of the 595 known hairpins were found. 68 of these are the same as test set 1, but 226 of the hairpins in the 517 sequences that were not family members of the six search model hairpins were also found. This indicates some ability to generalize, although the GA's fitness function was designed to make this number as large as possible (without getting lots of false positives).

In test set 3, 87 new hits occurred using the full length sequences. The additional sequence portions contain known hairpin structures that were not used by the GA to select the search model. Finally, the completely independent test set 4 revealed 92 hits on 561 sequences. This number is remarkably similar to the 87 hits gained on the independent portions of the 595 sequences in test set 3. Possibly, the hairpin structures within a family are no more similar than the hairpin structures from different families.

V. Conclusion

Truly generic non-coding RNA search models based on finding reasonable hairpin structures with acceptable false alarm rates are not achievable, as has been shown by Rivas and Eddy [16]. Since six out of sixteen possible combinations of two sequence positions can result in a Watson-Crick or a wobble base pair, there is a rather high probability that a random sequence will have two contiguous sub-sequences of lengths commensurate with those observed in known ncRNA molecules that are capable of forming nested base pairs and a distance between them that is a reasonable loop length. The alternative of forming family-specific models has the problem of requiring very large amounts of computation resources for search and will not find members of novel ncRNA families.

We have shown that it may be possible to find quite a few ncRNA genes without using a model targeted at a specific ncRNA family. This model requires far less computation time when compared to running the full set of family-specific ncRNA models found in Rfam since it only needs a single pass of the database. More importantly, it has been shown that some members of families that do not form any part of the search model are found. As a result, it is plausible that members of yet unknown ncRNA families may be found. We believe that this method of attempting to find a compromise between fully family specific and fully generic search algorithms may be potentially useful.

Reasons why this methodology might work is that not all feasible hairpin structures appear equally often in nature. It is known from laboratory thermodynamic studies that free energy changes in forming hairpins depends heavily on the specific nucleotides in the stem and loop end positions [18-22]. Also, some loop lengths are statistically more common than others. It is also clear that many ncRNA families are the result of gene duplication and subsequent differentiation. Novel ncRNA families may have primary or secondary sequence similarity with known families as a result.

The very preliminary nature of these results should, however, be stressed. The sixty four hairpin structures included in the list of possible structures in a search model were chosen randomly and the Rfam database contains many more structures than were in the list. The list size was small because a significant amount of non-automated human effort went into compiling this list. The testing is also very limited at this point. Only a single search model was tested due to time constraints. Ideally, the GA would be run many times and each resulting search model tested. The real test of this idea would be to run a large-scale database search using the search model and find and verify one or more ncRNA genes from novel families.

References

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

[2] S. Eddy and R. Durbin, "RNA Sequence Analysis Using Covariance Models," *Nucleic Acids Research*, Vol. 22, pp. 2079-2088, 1995.

[3] P. Gardner, J. Daub, J. Tate, E. Nawrocki, D. Kolbe, S. Lindgreen, A. Wilkinson, R. Finn, S. Griffiths-Jones, S. Eddy, and A. Bateman, "Rfam: Updates to the RNA Families Database," *Nucleic Acids Research*, Vol. 37, pp. D136-D140, 2009.

[4] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. Eddy, and A. Bateman, "Rfam: Annotating Non-coding RNAs in Complete Genomes," *Nucleic Acids Research*, Vol. 33, pp. D121-D124, 2005.

[5] Rfam: RNA families database of alignments and covariance models, version 9.1 (December 2008). http://rfam.janelia.org.

[6] S. Eddy, *Infernal user's guide*, version 1.0.2, Online at http://infernal.janelia.org, 2009.

[7] E. Nawrocki, D. Kolbe, and S. Eddy, "Infernal 1.0: Inference of RNA alignments," *Bioinformatics*, Vol. 25, pp. 1335-1337, 2009.

[8] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, et al. "Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology," *Comp. Appl. Bioscience*, Vol. 12, pp. 327-345, 1996.

[9] E. Nawrocki and S. Eddy, "Query-Dependent Banding (QDB) for Faster RNA Similarity Searches," *PLoS Computational Biology*, Vol. 3, e56, 2007.

[10] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, Vol. 205, No. 3, pp. 403-410, 1990.

[11] S. Altshcul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research*, Vol. 25, pp. 3389-3402, 1997.

[12] S. Eddy, "Hidden Markov Models," *Current Opinion Structural Biology*, Vol. 6, pp. 361-365, 1996.

[13] Z. Weinberg and W. Ruzzo, "Faster Genome Annotation of Non-coding RNA Families Without Loss of Accuracy," *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pp. 243-251, 2004.

[14] C. Burge and S. Karlin, "Finding the Genes in Genomic DNA," *Current Opinion in Structural Biology*, Vol. 8, pp. 346-354, 1998.

[15] T. Cech, J. Atkins, and R. Gesteland, *The RNA World*, Cold Spring Harbor Laboratory Press, 2006.

[16] E. Rivas and S. Eddy, "Secondary Structure Alone is Generally Not Statistically Significant for the Detection of Noncoding RNAs," *Bioinformatics*, Vol. 6, pp. 583-605, 2000.

[17] J. Smith, "RNA Search with Decision Trees and Partial Covariance Models," *IEEE Transactions on Computational Biology and Bioinformatics*, Vol. 6, pp. 517-527, 2009.

[18] T. Dale, R. Smith, and M. Serra, "A Test of the Model to Predict Unusually Stable RNA Hairpin Loop Stability," *RNA*, Vol. 6, pp. 608-615, 2000.

[19] M. Serra, M. Little, T. Axenson, C. Schadt, and D. Turner, "RNA Hairpin Loop Stability Depends on Closing Base Pair," *Nucleic Acids Research*, Vol. 21, pp. 3845-3849, 1993.

[20] M. Serra, T. Axenson, and D. Turner, "A Model for the Stabilities of RNA Hairpins Based on a Study of the Sequence Dependence of Stability for Hairpins with Six Nucleotides," *Biochemistry*, Vol. 33, pp. 14289-14296, 1994.

[21] R. Giese, K. Beschart, T. Dale, C. Riley, C. Rowan, K. Sprouse, and M. Serra, "Stability of RNA Hairpins Closed by Wobble Base Pairs," *Biochemistry*, Vol. 37, pp. 1094-1100, 1998.

[22] S. Freier, R. Kierzek, J. Jaeger, N. Sugimoto, M. Caruthers, T. Neilson, and D. Turner, "Improved Free-energy Parameters for Predictions of RNA Duplex Stability," *Proceedings of the National Academy of Sciences USA*, Vol. 83, pp. 9373-9377, 1986.