7-1-2006

A Genetic Algorithms Approach to Non-coding RNA Gene Searches

Jennifer A. Smith
Boise State University

This document was originally published by IEEE in *2006 IEEE Mountain Workshop on Adaptive and Learning Systems*. Copyright restrictions may apply.
DOI: 10.1109/SMCALS.2006.250691

A Genetic Algorithms Approach to Non-coding RNA Gene Searches

S. F. Smith
Department of Electrical and Computer Engineering
Boise State University
Boise, Idaho 83725-2075 USA
sfsmith@boisestate.edu
Phone: +1-208-426-5743, Fax: +1-208-426-2470

***Abstract*-A genetic algorithm is proposed as an alternative to the traditional linear programming method for scoring covariance models in non-coding RNA (ncRNA) gene searches. The standard method is guaranteed to find the best score, but it is too slow for general use. The observation that most of the search space investigated by the linear programming method does not even remotely resemble any observed sequence in real sequence data can be used to motivate the use of genetic algorithms (GAs) to quickly reject regions of the search space. A search space with many local minima makes gradient decent an unattractive alternative. It is shown that a fixed-length representation for alignment of two sequences taken from the protein threading literature can be adapted for use with covariance models.**

I. Introduction

The search for genes associated with functional non-coding RNA (ncRNA) molecules requires an algorithm that recognizes the conservation of base pairs in the RNA molecule even when there is little conservation in the primary sequence. Homology search algorithms such as BLAST [1], FASTA [2], Smith-Waterman [3], and hidden Markov models [4], can not accommodate these long range interactions. The hidden Markov model (HMM) approach can be extended from the use of a regular grammar which can not describe base pairing to a context-free grammar [5] which can. Such an extended model is often referred to as a covariance model (CM).

While a CM has the power to find ncRNA genes that an HMM can not, the computational burden of a CM is to great for it to be of use in most circumstances. DNA sequence databases are normally pre-filtered to find regions where a finding a gene of a given ncRNA family appears more probable. One method for doing this is to ignore base pairing and to start with an HMM constructed from a multiple alignment of the know members of ncRNA family [6]. This method risks losing sensitivity when the evolutionary distance between the known family members and the true unknown family member is large. A more recent method shows how to construct an HMM and choose a threshold such that no sensitivity is lost [7]. Even with this advance, the CM operating on the reduced database is still too slow for most purposes.

Interest in searching for ncRNA genes has increased in recent years as it has become increasingly apparent that many catalytic and regulatory functions depend directly on RNA molecules that are not translated into protein [8]. These ncRNA molecules may work either in isolation or as part of a complex of ncRNA and protein molecules. Examples of functional RNA participation include telomerase [9], small nucleolar RNA (snoRNA) [10], transfer RNA (tRNA) [11], and microRNA [12]. New ncRNA classes and families are being found at a very rapid rate.

Covariance models are composed of a binary tree of nodes associated with the consensus structure of the ncRNA family. Each of the nodes contains between one and six internal states. When scoring a covariance model against a database sequence, each state of the model is evaluated for every combination of starting location within the database sequence and subsequence length beyond the starting location. In order to make the calculation feasible, subsequence lengths searched are limited to an upper bound chosen by the user. This upper limit needs to be at least as long as the consensus sequence and is normally chosen to be considerably longer than the longest known ncRNA family member. Computational time required to score the model is proportional to this choice of subsequence length upper limit.

Investigation of the subsequence lengths actually observed in DNA databases indicates that deviations in subsequence length for true positives from the subsequence length obtained when fitting the consensus sequence are generally very small. This implies that a significant acceleration of the CM model may be possible if the solution search focuses in the region of small subsequence length deviations. It is difficult and time consuming for an expert to determine the upper and lower bounds that should be applied to each state evaluation. This would also remove length outliers entirely from consideration. An algorithm which can avoid getting trapped in the local minima of the score function, yet focus its search in the most likely region of small subsequence length deviations could speed the

1-4244-0166-6/06/$20.00 ©2006 IEEE

search considerably. A genetic algorithm (GA) is a good candidate for this task.

The paper is organized as follows. A short introduction to covariance models is given in section II. Section III investigates the subsequence length usage of two ncRNA families from different classes of ncRNA. The method for representing the alignment of the query sequence to the CM along with mutation operators and choices of initial population is described in section IV. Concluding remarks appear in section V.

II. Covariance Models

Covariance models can be estimated from a multiple alignment of sequences from a family of ncRNAs. The alignment needs to be annotated with structure information showing the intermolecular base pairing between nucleotides of the single-stranded RNA. The method does not model pseudoknots in the consensus structure, so some of the base pairing information in pseudoknotted ncRNA families is lost. A more complete description of covariance models may be found in [13].

A. Model Nodes from Consensus Sequence and Structure

Each emitting node of the CM is associated with either a consensus base pair or a consensus unpaired base in the structure-annotated multiple alignment of the ncRNA family. Figure 1 shows an example with five organisms. Alignment columns with few non-gap symbols are assigned a "." in the consensus structure and sequence and are not associated with any node in the CM. The ">" and "<" symbols indicate two consensus columns which tend to base pair, where the ">" is the nucleotide closer to the 5' end (the left base) and "<" is the nucleotide closer to the 3' end (the right base). Even though these symbols are not indexed, it is always possible to tell which go together since pseudoknots are not allowed. Column 4 and column 7 are base paired in the figure since there are no intervening base pair symbols. Columns 3 and 8 are base paired since they enclose a set of base pairs with no unpaired base paring symbols. The "-" symbol indicates that a column is associated with a position that does not base pair. The consensus structure for the family is determined either by experiment or RNA secondary structure prediction algorithms such as [14].

```
Organism 1      AUGG.ACCAAG.GUCAGACU
Organism 2      CUGAACUCCAGCGUCCGACU
Organism 3      CGG..GUCCCG.GA.AUU..
Organism 4      C.GAACUCG.G.GUCAGACU
Organism 5      CUCA.UUGUAG..UUA.ACU
Consensus:
Structure       ->>>.-<<-<-.>>---<<-
Sequence        CUGA.CUCCAG.GUCAGACU
```

Fig. 1. A structure-annotated multiple alignment.

The binary tree of the CM nodes for the model generated from the multiple alignment in Figure 1 is shown in Figure 2. Start (S), bifurcation (B), and end (E) nodes do not emit any symbols and are used only to form the tree structure for the three types of emitting nodes. The pair (P) nodes are associated with two paired consensus columns of the multiple alignment. Columns 4 and 7 in Figure 1 become the P node with index 8 in Figure 2. The P8 node is labeled with "AU" in Figure 2 to indicate that the consensus symbols of the pair are A on the left and U on the right. The left (L) and right (R) nodes indicate that an unpaired consensus symbol is to the left or right of the consensus subsequence represented by the nodes below it in the tree. For instance, the L1 node is associated with the first column of the multiple alignment and is labeled with a C since this is the consensus symbol at that position. There are situations where either an L or an R node could be used and by convention the L node is always used in these cases.

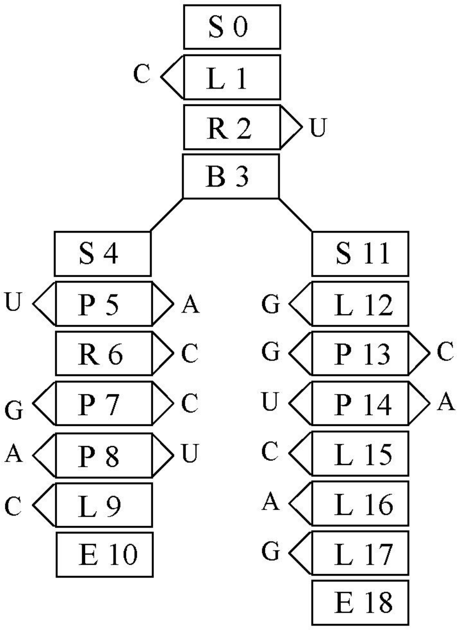

Fig. 2. Covariance model tree associated with structure-annotated multiple alignment.

The covariance model is a statistical model, so each of the nodes is associated with a set of probabilities for emitting each of the symbols. There are four emission probabilities associated with each of A, C, G, and U for the single emission L and R nodes and sixteen probabilities for each possible pair of emitted nucleotides in the P node. The node labels are simply the symbol or pair of symbols with the highest probability. The emitting nodes also allow for the possibility that the consensus symbol is omitted and the possibility that one or more symbols is inserted between it and the symbols of its children. The internal structure of the nodes which handles insertions and deletions relative to the consensus structure is described in Subsection B below.

Figure 3 shows how the consensus sequence and structure of the ncRNA family can be drawn in the form of a secondary structure diagram. The CM needs a bifurcation since the consensus secondary structure has two stems. The structure to the upper left of the bifurcation is captured in the left branch of the CM tree and the to lower right in the right branch of the tree. The exception to this is that the first and last columns of the multiple alignment are in the main stem of the CM tree since they are outside of all the P nodes. The two positions circled with the dashed line in Figure 3 are associated with the P8 node of the covariance model. The first nucleotide (5' end) is the consensus C symbol associated with the L1 node and the last nucleotide (3' end) is the consensus U symbol associated with the R2 node.

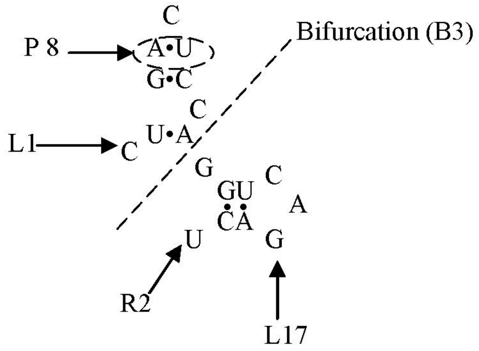

Fig. 3. Secondary structure diagram of consensus sequence.

B. Internal States of Model Nodes

The need to allow consensus symbols to be deleted and non-consensus symbols to be inserted is handled by the internal state structure of each node. The nodes consist of two possible tiers of states. The 1st (upper) tier handles the description of the consensus and the possibility that it is deleted. The 2nd (lower) tier handles the possibility that additional symbols are emitted between the consensus symbols of the node and the consensus symbols of a child node.

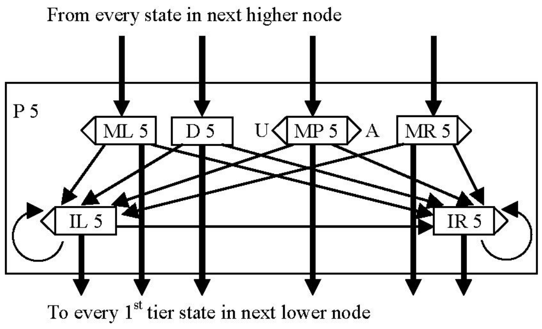
From every state in next higher node
P 5
ML 5
D 5
U
MP 5
A
MR 5
IL 5
IR 5
To every 1st tier state in next lower node

Fig. 4. Internal states of a P-type node.

Figure 4 shows the internal state structure of the most complicated type of the node, the P node. The upper tier includes a match pair (MP) state which is the consensus describing part of the node. It emits the sixteen possible combinations of symbols each with its own probability. If the database sequence does not contain the consensus pair at all, the delete (D) state is visited. If only the left nucleotide of the pair is present, the match left (ML) state is visited and this state emits each of the four possible symbols with its own probability. The match right (MR) state is used when only the right half of the consensus pair is present. The insert left (IL) and insert right (IR) states are used to add non-consensus symbols inside of the consensus pair, but immediately adjacent to the pair. The self loops on the IL and IR states allow more than one insertion. Figure 4 is drawn with "UA" next to the MP5 state since this P node is specifically the P5 node from the CM model tree. All nodes of a given type are structurally the same, only the probabilities associated with emissions and transitions between states differ. The thick lines in the figure indicate that there are multiple transitions possible in or out of a state coming from or going to states in another node.

C. Covariance Model Scoring and Subsequence Lengths

Every state in the CM is evaluated for every possible subsequence of the database sequence. In order to keep the computation tractable, only subsequences of length less than or equal to a maximum length D are computed. The score at each node ν is denoted $\gamma(\nu, j, d)$ and depends on the scores of the node's child states, the start position of the subsequence j and the length of the subsequence d. If insertions and deletions were not allowed relative to the consensus sequence, there would be only one state in each node and the only value of d needed for evaluation of that state would be the length of the consensus sequence represented by the node and all of the tree below the node. So, deviations in d from the consensus length at the state are due to net accumulated insertions and deletions at the state and all nodes below. The computational complexity of scoring the model is proportional to LD where L is the length of the database sequence and D is the maximum allowed subsequence length. The choice of D is traditionally the same for all states and must be chosen to be at least as large as the length of the consensus sequence. To this is added a guess as to the maximum number of insertions net of deletions that might occur in unknown family members. If the number of d values searched by each state could be reduced, then the computational complexity could be reduced by a factor equal to the ratio of D to the average number of d evaluations in each of the states.

III. Observed Subsequence Lengths

Two ncRNA families from different classes of ncRNA will be examined to show that large deviations from consensus subsequence length are rare for true positives in real DNA sequence data. The two families are taken from the Rfam database [6] and investigation of other families in the database indicates that the conclusions drawn from these two families is representative of most Rfam database families.

The first example is the miR-9 microRNA family (Rfam accession number RF00237) [12, 15-17]. There are five of these sequences taken from literature sources which form the seed population on which the CM is based. Searching a large DNA database using HMM pre-filtering and then using the CM yields another 18 putative family members. The average length of these sequences is 61 nucleotides. A condensed structure tree of the miR-9 CM is shown in Figure 5. It can be seen that the "tree" has no branches since this microRNA precursor (like microRNA precursors in general) has only a single stem structure. There are more than 45 microRNA families out of the approximately 500 ncRNA families in Rfam, so this example is not an atypical case. The compressed tree representation in Figure 5 omits showing S, B, and E nodes and groups multiple nodes of the same type into a single node. For example, the top node of Figure 5 implies that there are three L nodes at the top of the model and they have node numbers 1, 2, and 3 in the CM model file downloaded from the Rfam site.

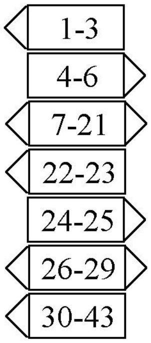

Fig. 5. Organization of the miR-9 (PF00237) model.

Table I shows the usage of subsequence lengths for each of the 23 family members for all states in the model listed by the number of the node that contains them. The high node numbers in this single-branch tree are near the end node and represent rather short consensus subsequences and the low node numbers are near the root start node and represent subsequences closer to the full consensus sequence. From the table we see that the first three L nodes above the E node (nodes 41, 42, and 43) have a symbol present in each of the 23 family sequences (no "." in the multiple alignment column). So, the consensus subsequence lengths are 1, 2, and 3 and the actual subsequence lengths are 1, 2, and 3 for all family members for nodes 43, 42, and 41 respectively. This results in the entry 23 under the 0 deviation column of Table I.

TABLE I
SUBSEQUENCE LENGTH DEVIATIONS FOR MIR-9 FAMILY

Nodes	-2	-1	0	+1	+2	+3	+4
43-41			23				
40		11	12				
39-37	7	4	11	1			
36-34	8	3	11	1			
33-29	8	3	11		1		
28-27	8	3	11			1	
26-1	8	3	11				1

The standard CM subsequence length upper limit *D* for this model used by Rfam is 100. Therefore the standard CM solution method searches length deviations between -1 and +99 for node 43, between -2 and +98 for node 42, and between -3 and +97 for node 41. The numbers in Table I indicate that the actual subsequence usage clusters near a deviation of 0, so large deviations are highly improbable. It would be possible to try to have a state-dependent upper and lower length cutoff, but this poses at least two problems. The first is that a large amount of effort is involved in determining good cutoffs for every state in every model. More importantly, it would be necessary to abandon all hope off finding outliers. In Table I we see that one true family member of miR-9 exhibits significantly more insert activity than the others (all of the 1 entries on the right side of the table are from a single sequence). If this sequence had not been in the initial model-building set, then we would be tempted to set the subsequence length cutoffs at -2 and 0 deviation for all states. The outlier sequence then would likely not be found. A better solution would be to not rule out rare subsequence lengths entirely, but instead just spend much less effort on them.

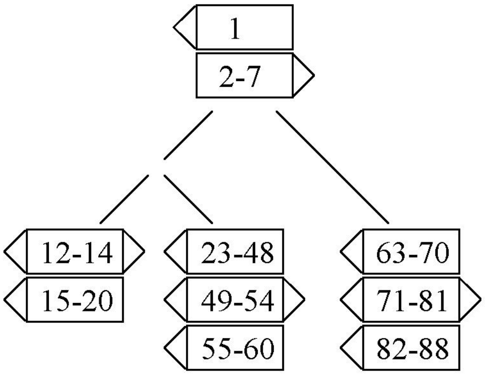

Fig. 6. Organization of the RyeB (PF00111) model.

The second example is the RyeB small RNA (Rfam accession number RF00111) [18]. Five of these sequences form the seed family used to train the CM and ten more were found using the CM in database search. The average sequence length of RyeB ncRNA family members is 100.

Figure 6 shows the structure of the CM for the family. This model has three stem structures, so the model tree exhibits branching (unlike the miR-9 model or microRNA precursor models in general). The RyeB model is part of a class of many small RNA models with similar branching patterns.

Table II shows the subsequence length usage of the RyeB example. Even though the model is quite different, the usage pattern is similar with most family members clustered very near 0 length deviation. The standard D value used for the RyeB model is 150, so nodes adjacent to E nodes (nodes 20, 60, and 88) will search subsequences with deviations between -1 and +149 even though in all cases the actual deviation is seen to be 0. Examination of other ncRNA family models leads to the conclusion that this clustering about 0 deviation is a general phenomenon and not limited to the two examples given or to the two classes of models to which they belong.

TABLE II
SUBSEQUENCE LENGTH DEVIATIONS FOR RYEB FAMILY

Nodes	-2	-1	0	+1	+2	+3	+4
88-81			15				
80-63			12	3			
60-58			15				
57		1	14				
56-24	1		14				
23	1	1	13				
20			15				
19-15			14	1			
14-12			11	14			
7-1		2	10		3		

IV. Representation For Model Alignement

A fixed-length representation for the alignment of a two sequences was presented in [19] for use in protein sequence three-dimensional structure prediction using threading. Here it is shown that this representation can be adapted for use in covariance models. Methods for choosing an initial population that focuses the search on regions of likely solutions and mutation operators are also presented.

The representation uses a string of non-negative integers of length equal to the consensus sequence of the CM. If a position in the string holds the integer 0, then the consensus nucleotide does not exist in the query sequence (a delete). If a position holds the integer 1, then a query symbol is to be matched to the CM node. For integers greater than 1, a query symbol is being matched, and additional query sequence symbols are inserted to the right. The number of inserted symbols is one less than the integer. In the case of L and R nodes, an integer larger than 0 implies that the ML or MR state of the node is visited. In the case of P nodes, two positions in the string need to be larger than 0 to visit the MP state. The ML or MR state is visited if only one of the two P node positions has an integer greater than 0.

In the protein threading case, we need to make sure that the sum of the integer values is equal to the length of the query sequence. The equivalent to the query sequence length in covariance models is the d value for the root S state score. In the linear programming approach, we want to search over all values of d between 1 and D. In the genetic algorithms approach, we can search all subsequence lengths simultaneously, so we do not need to constrain the sum of the integer values in the representation. This has the additional advantage of not needing to specify a cut-off length D in advance as is necessary for the linear programming method. For protein threading, it is necessary to choose a pair of representation positions and change the two values such that the sum of the changes is zero when mutating. Crossover is even more difficult since the child strings must be corrected to keep the integer sum constant. Neither of these problems is an issue in the application to a CM.

The analysis of Section III showed that most of the observed deviations from the consensus subsequence length are small. As a result, we expect the ultimate solution to be an individual with a representation not too far from the integer string with all entries equal to 1. To obtain a faster solution, it is recommended that the initial population contain at least one individual that is exactly the string of all 1 and that many other initial individuals be limited mutations of the string of all 1.

It is further noted from inspection of the ncRNA family members that inserts and deletes relative to the consensus occur in groups. A mutation of a single position in the integer string already takes care of generating runs of inserts at a given sequence position. To allow autocorrelation in deleted positions, mutations to 0 should be applied to a contiguous range of integer string positions. Both the statistical distribution of the number of 0 positions and the non-zero point mutation values should be estimated from the multiple alignment data.

In addition to autocorrelation in deletions and insertions across positions, the probability of insertion and deletion at each position varies greatly. The observed frequency of insertions and deletions in the ncRNA family multiple alignment can be used to generate a position-specific mutation operation. The autocorrelated and position-specific mutation operator should be used for both initial population generation and for subsequent evolution.

V. Conclusions

We have seen that there is a large amount of probabilistic order to the search space examined by covariance models used for ncRNA gene finding. Linear programming methods enumerate the entire search space and hence are very slow (but are guaranteed to find the best solution). Many of the solutions evaluated by linear programming are of exceedingly small probability. However, putting tight constraints on the search space is difficult and risks missing a few outlier true solutions. The proposed GA solution method does not limit the potential search space (in fact it increases it by not introducing a hard upper limit on subsequence size). The GA solution gains performance by probabilistically focusing on the most probable portions of the search space leading to an algorithm which will find nearly the best score in many fewer solution evaluations.

It should be noted that the output of the GA search can also be used as a pre-filter to a full linear programming CM search. Since the GA may produce a suboptimal score, the threshold for accepting a subsequence as a family member of the ncRNA family could be set lower than usual and the family members found could be rescored using the full CM to get exact scores. The full CM would run on a very small subset of the original data and therefore the computational requirements may be tenable.

Further research in this area clearly includes quantifying the tradeoffs between computation time and score accuracy. It is also hoped to investigate searching over both position *j* and subsequence length *d* simultaneously (rather than just *d* as presented in this paper). This could have the advantage of rejecting segments of the query sequence with little hope of scoring well in favor of increased effort in proximity of high scoring positions.

Acknowledgment

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

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, 215, pp. 403-410, 1990.

[2] W. Pearson and D. Lipman, "Improved Tools for Biological Sequence Comparison," *Proceedings of the National Academy of Sciences*, 4, pp. 2444-2448, 1988.

[3] T. Smith and M. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, 147, pp. 195-197, 1981.

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

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

[6] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. Eddy, "Rfam: An RNA Family Database," *Nucleic Acids Research*, 31, pp. 439-441, 2003.

[7] 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.

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

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

[10] T. Kiss, "Small Nucleolar RNAs: an Abundant Group of Noncoding RNAs with Diverse Cellular Functions," *Cell*, 109, pp. 145-148, 2002.

[11] Y. Hou, "The Tertiary Structure of tRNA and the Development of the Genetic Code," *Trends in Biochemical Science*, 18, pp. 362-364, 1993.

[12] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of Novel Genes Coding for Small Expressed RNAs," *Science*, 294, pp. 853-858, 2001.

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

[14] M. Zucker, "Computer Prediction of RNA Structure," *Methods in Enzymology*, 180, pp. 262-288, 1989.

[15] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, "Identification of Tissue-specific microRNAs from Mouse," *Current Biology*, 12, pp. 735-739, 2002.

[16] N. Lau, L. Lim, E. Weinstein, and D. Bartel, "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans," *Science*, 294, pp. 858-862, 2001.

[17] L. Sempere, N. Sokol, E. Dubrovsky, E. Berger, and V. Ambros, "Temporal Regulation of microRNA Expression in Drosophila melanogaster Mediated by Hormonal Signals and Broad-complex Gene Activity," *Developmental Biology*, 259, pp. 9-18, 2003.

[18] K. Wassarman, F. Repoila, C. Rosenow, G. Storz, and S. Gottesman, "Identification of Novel Small RNAs Using Comparative Genomics and Microarrays," *Genes and Development*, 15, pp. 1637-1651, 2001.

[19] J. Yadgari, A. Amir, and R. Unger, "Genetic Threading," Constrains 6, pp. 271-292, 2001.