Chris Moore
Biochem 218
Final Paper

Computational Methods of Identifying Novel sRNA Genes in *E. coli*

Introduction:

Despite the fact that the existence of sRNAs has been known for some time, most gene identification methods and programs are geared to find only genes that code for proteins, and are incapable of finding genes whose functional form is the RNA itself. It is only recently, with the explosion of identified sRNA genes and the belief that many more exist, that concerted efforts are being made to develop methods that can identify all of the sRNA genes in the genome sequences available, much as has been done (to a certain extent) with protein coding gene-finding algorithms. However, developing accurate computational methods of whole genome scale sRNA gene finding has proven difficult, for a variety of reasons (Eddy, 2001).

Small RNAs are defined as genes whose functional product is the transcribed RNA, and which do not code for proteins. Small RNAs come in many different flavors, with a variety of functions and characteristics (see Box 1). This makes it extremely difficult to create an algorithm that can identify them all from raw sequences data, especially as not all of the characteristics of each group are known yet (Chen *et al.*, 2002).

Box 1: Abbreviations for different classes of non-coding RNA (ncRNA or sRNA)

- fRNA Functional RNA — essentially synonymous with non-coding RNA
- miRNA MicroRNA — putative translational regulatory gene family
- ncRNA Non-coding RNA — all RNAs other than mRNA
- rRNA — ribosomal RNA
- siRNA mall interfering RNA — active molecules in RNA interference
- snRNA mall nuclear RNA — includes spliceosomal RNAs
- snmRNA Small non-mRNA — essentially synonymous with small ncRNAs
- snoRNA Small nucleolar RNA — most known snoRNAs are involved in rRNA modification
- stRNA Small temporal RNA — for example, *lin-4* and *let-7* in *Caenorhabditis elegans*
- tRNA Transfer RNA

Eddy, 2001

Small RNAs are also relatively difficult to identify *in silico* due to the fact that their sequences have no obvious inherent statistical biases that can be exploited, as do exon coding genes (*i.e.*, conservation of the ORF) (Rivas and Eddy, 2000). The lack of known and recognizable features, coupled with the lack of obvious sequence biases, make

finding sRNA gene a very difficult challenge to bioinformaticists, even in an organism as well studied as *E. coli*.

Despite these challenges, several groups have recently made an effort to produce methods of computationally identifying various classes of sRNA genes in *E. coli*. They have used a variety of methods, focusing on various characteristics recognized in the currently known sRNA sequences in *E. coli*. The three major computational methods of sRNA gene-finding used by the four groups discussed are listed in Box 2. All four groups used one or more of these methods in their algorithms.

Box 2: Three Methods of Computational sRNA Gene-finding

1. Prediction of RNA transcriptional signals
   a) promoters
   b) terminators
   c) lack of potential small ORFs
2. Sequence content statistics
   a) secondary structure stability
   b) base composition
3. Comparative genome analysis
   a) conservation of sRNA sequence between related species
   b) conservation of secondary structure by compensatory base changes

Methods of identifying sRNA genes:

In one approach to identifying sRNA sequence in *E. coli*, Wasserman *et al.* (2001) used conservation of intergenic regions (IGs) between *E. coli*, *Salmonella pneumonia*, and *Klebsiella pneumonia* to generate a list of putative sRNAs. When examining the 13 small RNAs known in *E. coli* at the time of their experiments, they noticed that these genes were well conserved in closely related bacteria. While typical ORF genes showed <70% conservation, the sRNA genes showed >85% conservation between *Salmonella* and *E. coli*. They hypothesized, based on this difference in relative conservation and on positive tests of conservation of random noncoding regions between the same organisms, that extended conservation within intergenic regions was statistically significant enough to use for predicting novel sRNA genes. To this end, they compared all IG regions 180 nucleotides or greater in length to similar regions in *S. pneumonia* and *K. pneumonia*, two closely related bacterial species. After finding 1097 regions of high homology, regions containing tRNAs, rRNAs, repetitive sequences, putative ORF regulatory regions, and previously identified ORF promoters or 5' UTRs were removed. Evidence of putative promoters, rho-independent terminators, and stem loops were considered especially indicative of potential sRNA genes. Using this computational analysis and information from high-density oligonucleotide arrays, including higher expression levels and expression of both strands, their list was narrowed down to 59 candidates.

To determine which of their candidates actually coded for sRNAs, the standard method of transcript detection by Northern blotting was performed. Of the 59 putative transcripts, 23 were detected as small transcripts. For these sequences, the conserved

blocks of sequence from *K. pneumonia*, *Salmonella*, and *Yersinia pestis* were selected and aligned by hand for evidence of promoters and terminators. While doing this, it was noticed that six out of the 23 transcripts showed a pattern of conservation corresponding to a short ORF (higher variation in positions that could be the third nucleotide of a codon), and were determined to have potential ribosome-binding sites. These were eliminated as putative short ORF genes. The remaining 17 transcripts showed no evidence of translation potential, and were tentatively designated as novel sRNA genes.

After the completion of their experiments, Wasserman *et al.* (2001) conclude that, while they were able to identify IG regions containing novel RNA sequences at a fairly well (~30% of the selected candidates encoded novel small transcripts, and ~29% of the selected candidates encoded novel sRNA genes), a high level of sequence conservation is not by itself sufficient to indicate small RNA genes. Most of the highly conserved sequences found corresponded to the regulatory regions and UTRs of flanking regions, and even an attempt to identify and remove these sequences was insufficient and needed to be complemented by a high density oligonucleotide array. Also, not all sRNAs are conserved between even closely related species, and the sRNAs that are processed from mRNAs or encoded by an ORF antisense strand will be missed. Despite these inherent flaws, this computational method does show promise as a possibly supplemental way to initially identify a pool of sequences that can be examined by more sophisticated methods for sRNA genes, as it can be utilized in any organism for which the sequences of closely related species are available, and can find sRNAs that do not have a stable secondary structure. It can also be used to identify and group families of sRNA genes based on their pure sequence conservation between species.

In a second approach, Argaman *et al.* (2001) started their search for sRNA genes in *E. coli* by examining the current known sRNA sequences and developing a set of criteria that appeared to be similar for all of the known sequences.

1. located in "empty" intergenic intervals between annotated protein coding genes
2. conserved in some closely related species
3. transcriptional signals: promoter shortly upstream of terminator

Due to the limited number of known sRNAs, they used a heuristic approach instead of an automatic machine-learning approach. Using their developed criteria, they first compiled a list of all "empty" IG regions in *E. coli* using the Colibri database (http://genolist.pasteur.fr/Colibri/) and searched for the most commonly used transcriptional promoter and terminator sequences. These common signals were those corresponding to the major *E. coli* RNA polymerase sigma factor, σ^{70}, and to Rho-independent terminators, which are recognizable by specific sequences and structural features in the RNA. Their list of predicted signals was then narrowed down to those in which the predicted promoter and terminator signal pairs were 50-400 base pairs apart, approximately the lengths of the known sRNAs. These predicted sequences were then compared to the genome sequences of *Salmonella typhi*, *S. paratyphi*, and *S. typhimurium*. Those sequences showing significant conservation between all three

organisms (24 putative sequences) were further analyzed by Northern blotting to confirm the presence of small transcripts, and then assayed by primer extension analysis to determine the true ends of the transcripts. From the 24 putative sRNA genes, 23 were tested experimentally in K12 *E. coli* at different growth phases, and of these 14 were shown to encode novel small RNAs.

After determining the presence of small transcripts, the positions of these sequences in the *E. coli* genome were examined, along with flanking sequences. It was determined that most of the new sRNA genes were clustered in short IG regions of 600 base pairs or less. Also, it was seen that there were two major types of conservation patterns of sRNA genes with respect to their flanking sequences. For some of the transcripts, the entire surrounding region, including both flanking IG sequences and genes, was conserved between all of the bacteria examined, while for the rest, the surrounding IG sequences (and some genic sequences) were unconserved. This reveals that conserved sRNAs can potentially function in a variety of genomic contexts, and are not necessarily dependent on surrounding genes.

The results gained from these experiments show that, in organisms for which transcriptional signals are well characterized, such as *E. coli*, a search for these transcription signals in IG regions can be a powerful tool for finding potential small RNA genes, especially when coupled with comparative analysis with other organisms. Of course, not all sRNA genes will be conserved between the organisms studied, and as such there may be a difference in the false negative rate based on which other genomes are analyzed. This example does indicate that a rule based approach, based on previously identified sequences, can be useful, and as more sRNAs are identified, more characteristic can be identified and utilized as well, making it a more precise search method. However, it is already evident that not all sRNAs follow all of the same rules. Thus, the addition of certain rules or criteria to the search, while increasing its sensitivity for certain sRNAs, will likely also increase the false negative rate by not identifying sRNAs that do not share all of the same characteristics. When using a criteria-based system such as this, it is important to understand that only the small RNAs that fit into the specific class that is described by your chosen criteria will be found by your search algorithm, and others will be excluded. One possible way of allowing the use of a greater number and more specific criteria is to allow weighting of various combinations of criteria (necessitating the creation of new criteria), such that certain combinations seen in currently known examples result in raising the score, while other combinations not seen in currently known examples results in lowering the score. Of course, there is always the possibility that sRNAs containing the negative set of criteria do exist and just have not yet been found, and the program will of course not be able to find these.

In yet another approach to computationally identify sRNAs in *E. coli*, Carter, Dubchak, and Holbrook (2001) designed a machine learning approach that utilized neural networks and support vector machines to identify the common features present in known sRNA sequences to predict novel sRNA genes in IG regions. Their hypothesis was that, despite the currently apparent lack of statistical signals heralding sRNA genes, characteristic signals for the expression of sRNAs must exist and be distinguishable from

untranscribed sequences. In utilizing this approach, they focused on two aspects of sequence analysis: secondary structure stability and base composition. The secondary structure stability was analyzed by determining the free energy of folding within each RNA sequence window considered, and base composition was represented by the percentage of each nucleotide and dinucleotide pair in the same windows. After considering both sets of parameters separately, a final voting network was used to consider and weigh the results from both input networks and return a decision on whether or not the sequence was indicative of a sRNA gene. In their tests, they predicted approximately 370 novel sRNA genes in *E. coli*, and while jackknife testing experiments revealed 80-90% accuracy, they did not perform any biological assays to confirm their results.

In applying their networks to the IG regions of *E. coli*, they used a buffer zone of 50bp on either side of the ORF to account for 5' and 3' UTRs (an approach noticeably lacking in the other groups' analyses). While this was a good attempt in theory, in practice it would be better to know (or at least statistically approximate) the extent of the true transcripts, perhaps by looking for putative promoters and terminators as in Argaman *et al.* (2001). Also, this approach doesn't take into account the existence of operons, although whether or not sRNAs can be coded by the inter-ORF sequences of operons has not yet been determined. After extracting the shortened IG regions, they had to use these non-annotated IG sequences as negative examples of RNA genes in order to train their machine, while using the small number of known sRNAs as positive examples. Unfortunately, there are very likely some sRNA genes present in these non-annotated IG sequences. They assume that the RNA coding sequences make up only a small fraction of the non-annotated sequence, and are thus justifiable in being used, but they are still making an assumption. This assumption cannot be proved valid or invalid by their own methods, either, as they use their program to detect and remove putative sRNA sequences from their non-coding database and retrain on the purified database. This may lead to overtraining of their program, in that sRNA sequences missed by the first iteration will be continually considered as non-coding sequence, and continual training will remove any chance of these sRNAs being uncovered by the program. Of course, should these sequences be uncovered and confirmed by other methods, they can be removed from the non-coding database and the program can be retrained, and possibly uncover others of these types.

In utilizing secondary structure stability as one of their inputs for their network, Carter, Dubchak, and Holbrook (2001) disagree with the findings of Rivas and Eddy (2000), which state that the predicted stabilities of the secondary structures of most sRNAs are not sufficiently different from those of random sequences to be used as a criterion for identification. While Carter, Dubchak, and Holbrook agree that random sequences are not sufficiently different from sRNAs, they postulate that the non-coding sequences (that actually exist in *E. coli*, as opposed to purely random sequences, which have not been proved to exist) used in their analysis have their own sequence biases, which result in a statistically significant difference in secondary structure stability from that of sRNAs. They also state that the removal of secondary structure stability of putative sequences from their networks greatly reduces the algorithm's ability to

accurately identify novel sRNAs (based on a comparison with the ones they originally identified).

This method of sRNA identification has several key strengths. First, any predictions made by the machine can be checked for agreement between each of the different networks, potentially reducing the rate of false positive predictions. Second, each input parameter (the different sequence composition statistics or free energy of folding determinations), and potentially any new parameter introduced, can be individually assessed and weighted based on its relative importance in the prediction. This make the program very flexible, and potentially able to adapt to the differences between sRNA genes, making it a feasible approach to identify more than one type of sRNA based on different sets of criteria. And finally, this approach may be able to predict sRNAs in many different organisms, as it is not inherently organism-specific, and can be retrained for different genomes. Also, as more sequences (both of sRNAs and true non-coding IG regions) become available, further training and iterations of this approach will only become more accurate. Unfortunately, without any further analyses, it is not possible to comment on how well it truly identified novel sRNAs, and as such its success is still in question.

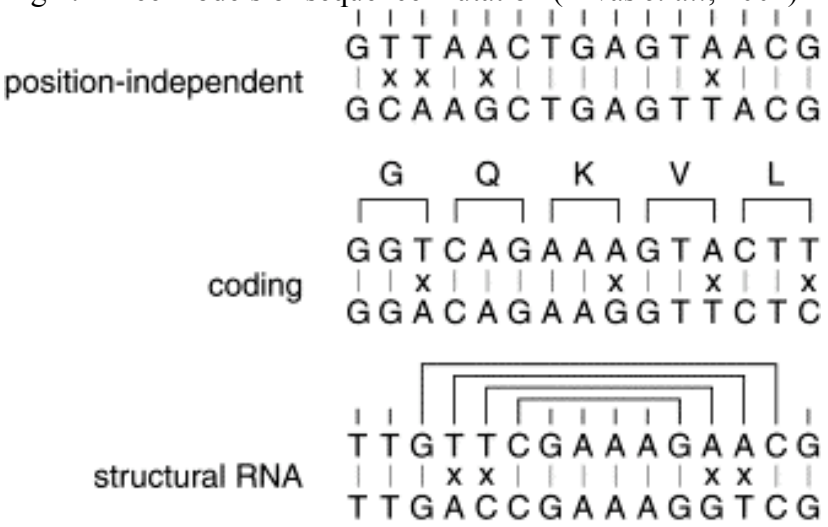
In the final approach to identifying sRNAs in *E. coli*, Rivas *et al.* (2001) focused on the conservation of secondary structure recognized to exist in some of the known sRNAs. They had previously determined that a low energy secondary structure in a sRNA was not statistically different enough to differentiate it from possible secondary structures in random sequences (Rivas and Eddy, 2000). However, they recognized that sequences within genomes are conserved for many different reasons, including coding sequences for exons and sequences that produce secondary structures in expressed RNAs, and as such the ways in which they are conserved differ as well. They used the idea that different sequences are conserved in different ways to design an algorithm that looks for distinctive patterns of mutation in conserved intergenic regions that correlate with a conservation of RNA secondary structure, and contrasted that pattern to patterns of mutation that reflect coding sequences or "random" sequences.

In this way they created a program (QRNA) that detects novel structural RNA genes (those that form a coherent and functional secondary structure *in vivo*) by means of a comparative sequence analysis algorithm. Their algorithm is composed of three stochastic "pair grammar" probabilistic models that each sequence in IG regions can be modeled as (see Fig 1).

RNA: mutation pattern consistent with structural RNA sequence, modeled by pair stochastic context-free grammar (SCFG)
COD: mutation pattern consistent with protein coding sequence, modeled by a pair Hidden Markov Model (HMM)
IND: independent pattern of mutation, the "null hypothesis", modeled by a pair HMM

In using this method, it is important to note that all three models need to be at the same evolutionary distance, or they may distinguish alignments on the level of conservation alone instead of on the different patterns of mutation (Rivas *et al.*, 2001).

Fig 1: Three models of sequence mutation (Rivas *et al.*, 2001)



This approach was applied to the IG regions of *Escherichia coli* in comparison to sequence data from four related enterobacterial genomes, those of *Salmonella typhi*, *S. paratyphi*, *S. enteriditis*, and *Kleisella pneumonia*. Each IG region of *E. coli* was BLASTed against the other genomes, and those alignments with an E value of less than 0.01, a length of 50 nucleotides or more, and an identity of 65 or greater were retained, resulting in 23,674 total pairwise alignments. Each of these was analyzed by QRNA to determine which sequences matched the RNA class of secondary structure conservation, resulting in 556 candidates (including all four positive controls of known sRNAs included in the test). Various tests (Rivas and Eddy, 2001) were conducted on QRNA's performance suggesting a sensitivity of 80%, and a specificity of about 85% (meaning that about 85% of the 556 candidate loci should correspond to sequences with true conservation of secondary structure). Several classes of sequences have conserved secondary structures but are not sRNA genes, including rho-independent terminators, rRNA spacers, and other *cis*-regulatory RNA structures; any loci showing evidence of these features were removed, leaving 275 candidate loci. 49 of these candidate loci were assayed for expression by Northern blotting, and 11 showed RNA transcripts of less than 400 nucleotides. However, despite the fact that several known RNAs are expressed only under certain conditions, they only used one set of growth conditions, and as such negative results do not necessarily indicate the absence of a sRNA gene.

The major strength of this type of analysis is that it doesn't require any organism-specific information, such as transcription promoter consensus sequences or terminator structures, and thus can be used on any genome in which there is sufficient comparative sequence data. Unfortunately, it also suffers from several weaknesses. The fact that it requires comparative sequences to identify putative sRNA genes means that it cannot be used on an independent genome for which no comparable sequence information is available. Also, it was shown that the results of the initial run of the program have a greater than 50% false positive rate, in the form of structurally conserved elements of mRNAs or other non-sRNA sequences. This means that any results must be further analyzed to remove these types of sequences, and the success of this removal depends on knowledge of the elements that correspond to translational machinery in the specific

organism, reducing the breadth of genomes on which it can be used. And finally, it is well known that not all sRNA genes have conserved secondary structure, or are conserved between genomes at all, and these sequences will not be identified by QRNA. This program demonstrates a method of finding only a very specific subset of sRNA genes, based on one major criterion, and as such, cannot be used as a general sRNA gene finding program.

Proposal of possible improvements:

In the preceding discussions of methods geared towards identifying novel sRNAs in the *E. coli* genome, it has become readily apparent that no one single method is capable of finding all of the different types of sRNAs, even in an organism as well studied as *E. coli*. This can be attributed both to our still incomplete understanding of sRNAs in general, as well as the fact that the different types of sRNAs all seem to have a fairly individual signature. Structural sRNAs maintain a secondary structure through compensatory mutations, while siRNAs and miRNAs must conserve complementary sequence with their targets. And the fact that not all sRNAs are conserved in all genomes reduces the power of comparative studies. Thus, it appears necessary to develop a group of computational methods that are each specifically tailored to a different class of small RNAs, such that while each program will only find a specific subset of sRNAs, it will be able to find them with a high fidelity.

There are also many possibilities of improvements and expansions to the existing programs previously outlined, such that a greater flexibility and accuracy can be imparted. For example, many endogenous siRNAs are transcribed from the antisense strand of ORFs (Eddy, 2001), and as such they are evolutionarily linked to the sequence they regulate, and will always maintain perfect complementarity. All efforts to identify sRNAs to date have focused on IG regions, and will consequently miss these siRNAs. It is possible to examine ORFs for putative internal promoters on the antisense strand, at least in *E. coli* where the σ^{70} consensus sequence is known. Also, as these sRNA sequences are most likely tightly controlled either temporally or environmentally, a better understanding of sequences that control temporal or environmental transcriptional responses would allow for better *in silico* sRNA gene-finding programs, as they could look not just for putative promoters and terminators, but also for putative control sequences. While not entirely a computational approach, another likely method of finding sRNAs that are present in ORFs, which all of the programs described would miss, would be, as in the Wasserman *et al.* (2001) paper, examination of high density oligonucleotide arrays for transcripts corresponding to the antisense strand of ORFs (in the majority of instances for which two different genes are not coded on each strand). Computational methods can be developed to quickly examine these arrays specifically searching for this evidence.

In Rhoades *et al.* (2002), they discuss using the sequence of previously identified miRNAs in *Arabidopsis* to identify genes that they could potentially regulate. This approach works due to the fact that the miRNAs (or at least the functional ~22nt piece cut from the hairpin precursor) have near-perfect complementarity to their targets. This suggests an approach for finding miRNAs *de novo*, namely by comparing intergenic

sequences to known or predicted mRNA sequences, looking for ~22nts of near-perfect complementarity. One problem with this is that miRNAs in other organisms aren't as perfectly complementary as they are in *Arabidopsis*, so how much mismatch and where it is acceptable must be determined. Zeng and Cullen (2003) have made an important start in this direction by determining that a specific miRNA in humans, miR-30, cannot completely discriminate between targets that differ by only a single nucleotide. However, certain point mutations in the targets *are* able to prevent suppression of translation by miR-30, suggesting that certain nucleotides are far more important to target recognition than are others. Further work may help determine if this is characteristic of all miRNAs, and, if so, if it can be determined computationally which nucleotides must stay complimentary, and which can be variable.

If phylogenetically conserved miRNAs have acquired multiple antisense targets, thereby inhibiting their subsequent evolutionary variation (Ambros *et al.*, 2003), then a program could be created that searches for conserved miRNAs in other species, comparing not only sequences between species, but also sequences within species to see if a putative conserved miRNA complements the same homologous genes between species. This would allow further definition of miRNA functions, and can be combined with secondary structure predictions and examinations of IG sequence upstream and downstream of the identified putative miRNA looking for complementarity within the IG region and stable secondary structures, as miRNAs are processed from RNA hairpins (Lee, Feinbaum, and Ambros, 1993).

In developing better computational methods of sRNA gene-finding, more must be understood about the biology of sRNAs, and of the organisms in which they are studied. In order to examine IG sequences, it is necessary to know where the actual intergenic regions are, instead of just using inter-ORF sequences. This will require further knowledge about promoters and terminators to allow computational identification, or a concerted biological assay to identify full length transcripts. Also, a program that can recognize more that just the σ^{70} transcription factor would be a good start, but more work needs to be done to learn about the different transcription factors to be able to recognize their consensus sequences. Also, as rho-dependent transcription terminators are poorly characterized and no consensus rho factor binding site is known (Lewin, 1994), it would be extremely difficult to identify any sRNAs that are transcribed in this manner without further knowledge.

Perhaps a way of identifying multiple types of sRNAs in a genome could make use of an expanded version of the neural network described in Carter, Dubchak, and Holbrook (2001), which can balance a greater variety of statistics than were used in the original iteration. As more sRNA sequences are identified, more characteristics, and more importantly, how those characteristics group in each of the defined families of small RNAs, can be used to build a type of scoring matrix that can identify multiple characteristics in a single sequence and score each characteristic in the context of the others. This would create not only a single yes or no score, but also a score that places the putative sRNA sequence into a previously existing family of small RNAs, facilitating faster biological identification by the appropriate methods. In the end, it is apparent that,

as of now, we do not know enough about small RNAs to develop such a program, and continued efforts must be made at biological and computational identification to increase our knowledge to the point where it is possible.

Works Cited

Ambrose, V, RC Lee, A Lavanway, PT Williams, D Jewell. MicroRNAs and Other Tiny Endogenous RNAs in *C. elegans*. *Current Biology*, 13:807-18. 2003.

Argaman, L, R Hershberg, J Vogel, G Bejerano, E Gerhart, H Wagner, H Margalit, S Altuvia. Novel small RNA-encoding genes in the intergenic regions of *E. coli*. *Current Biology*, 11:941-950. 2001.

Carter, RJ, I Dubchak, SR Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research*, 29:19:3928-38. 2001.

Chen, S, EA Lesnik, TA Hall, R Sampath, RH Griffey, DJ Ecker, LB Blyn. A bioinformatics approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 65:157-77. 2001.

Eddy, SR. Non-coding RNA Genes and the Modern RNA World. *Nature Reviews: Genetics*, 2:919-929. 2001.

Eddy, SR. Computational Genomics of Noncoding RNA Genes. *Cell*, 109:137-140. 2002.

Lee, RC, RL Feinbaum, and V Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75:843-54. 1993.

Lewin, B. In: Genes V. (Ed.), Control by RNA Structure: Termination and Antitermination. Oxford University Press, Oxford, pp. 461-2. 1994.

Rhoades, MW, BJ Reinhart, LP Lim, CB Burge, B Bartel, DP Bartel. Prediction of Plant MicroRNA Targets. *Cell*, 110:513-520. 2002.

Rivas, E and SR Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 17:391-7. 2000.

Rivas, E and SR Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8-27. 2001.

Rivas, E, RJ Klein, TA Jones, SR Eddy. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current Biology*, 11:1369-73. 2001.

Zeng, Y, and BR Cullen. Sequence requirements for micro RNA processing and functin in human cells. *RNA*, 9:112-123. 2003.

Appendix: Summary of the strengths and weaknesses of the computational methods of identifying sRNA genes in *E. coli*

Wasserman *et al.*, 2001

Strengths	Weaknesses
-Very productive in identifying IG regions that encode novel small transcripts (>30% of putative candidates encoded small transcripts) -Can be used for any organism with sequences of closely related organisms available	-High level of conservation between organisms is insufficient to indicate sRNA genes -Misses any sRNAs processed from mRNAs or encoded by an ORF antisense strand -Not all sRNAs are conserved between even closely related species

Carter, Dubchak, and Holbrook (2001)

Strengths	Weaknesses
-Predictions can be checked for agreement between the different networks -Each input parameter can be weighted based on its importance to the prediction -May be able to predict sRNAs in many different organisms after retraining -As more sRNAs are identified, more criteria can be included, increasing its usefulness	-Putative transcripts were not tested biologically -Potentially too few or ambiguous criteria -Possibility of overtraining program to not identify certain sRNAs

Argaman *et al.*, 2001

Strengths	Weaknesses
-In organisms for which transcription signals are well understood, it is a powerful biological principle based approach, rather than a purely statistical one -As more sRNAs are identified, more characteristics can be defined, making a characteristic driven program better able to recognize novel sequences	-Uses the consensus sequence of only the major sigma factor, missing any sRNAs associated with alternative sigma factors or stress-specific transcription factors -Uses only rho-independent terminator sequences, missing any rho-dependent terminators -Requires detailed knowledge of the organism's transcriptional machinery -Presence of multiple potential promoters and bidirectional terminators can confuse the prediction -Promoter signals are relatively weak

Rivas *et al.* (2001)

Strengths	Weaknesses
-Doesn't use any organism-specific info, so it can be used on any organism with sufficient comparative sequence data to identify candidate structural ncRNA loci	-Requires comparative sequence data, cannot be used on an independent genome -Detects nongenic sequences that show conserved RNA structure (false positives) -Not all ncRNA genes have conserved intramolecular structures, and will be missed by this program (false negatives)