Madeline Dinsdale
May 17, 2016
Computational Biology
Final Project Report

Identification and Characterization of pre-miRNA Candidates in the *C. mitchellii* Genome

Summary

MicroRNAs (miRNAs) play an important role in the regulation of many biological pathways. Identification of these miRNA sequences represent an important part of understanding genome function. Genomic sequencing analysis has shown that the vast majority of eukaryotic multicellular genomes are made up of non-coding and/or repetitive sequences. These regions, previously considered nonfunctional 'junk DNA,' actually perform important regulatory functions. Some of these non-coding regions are transcribed and the resulting RNA sequences fold to make complex structures. These small RNAs are never translated to a protein, however they perform other functions in the organism. One type of small RNA, microRNA (miRNA) play a role in gene regulation. Identification and characterization of these miRNAs within the genome provides valuable insight into the molecular mechanisms underlying the regulation of differential protein expression. In order to compile a library of candidate miRNA sequences for the C. mitchelli genome, a portion of the sequenced *Crotalus mitchellii* (Speckled rattlesnake) genome was analyzed for the presence of candidate pre-miRNA sequences based on secondary hairpin structure using the Ab initio web program miRNAFold (http://omictools.com/mirnafold-tool) and a program written in Python that plots the miRNAFold output data against several key miRNA parameters.

Data Formatting

The whole genome shotgun sequence of *C. mitchelli* assembled in 2013 was obtained from the NCBI genome database. Of this genome, which consists of 473,380 contigs, the 50 longest contigs were converted to text files and analyzed for possible miRNA sequences. Candidates were identified using miRNAFold, an Ab initio miRNA prediction tool. The output of this program contains the start and stop position on the contig, the size of the miRNA sequence identified, the minimum free energy of the structure, the RNA sequence and two visual representation of the stem-loop structure (see figure 1). The miRNAFold prediction parameters we set to sliding window size of 150 nucleotides, minimum hairpin size of 30 nucleotides, a minimum percentage verified features of 70%, and the species parameter set to all miRNAbase genomes (non-species specific). The result files containing all candidate miRNA data for each of the 50 contigs were manually converted to text files for further computational processing.

Program Steps

My program stores the text files containing the information for candidate miRNA information into a dictionary and then creates four line plots for each contig which map the miRNAs. The first part of my program was made in collaboration with Elaine Kushkowski (see citations in program). The first step of the program is to import a text file with a list of the contig numbers, which is used to open each contig text file (OpenFile function). Then, for each contig our program stripped the miRNAFold output files of all unnecessary information and stored all candidate miRNA information in a dictionary for each contig. The candidate miRNA dictionaries were then refined to exclude miRNAs that do not fall below a threshold free energy value, which

is set manually in the main function (lowest_freeEnergy function). The purpose of this function is to be able to exclude miRNAs that have a high free energy and are unstable. The refined candidate miRNA dictionaries where then stored in an outer dictionary by contig number (makeDictionary function). This gives us a dictionary of dictionaries where the miRNA sequences and associated information such as starting indices, ending indices, and free energy are stored according to contig number and starting index keys.

The second part of the program was done individually. For each contig in the dictionary, all candidate miRNAs with a free energy value below -60 kcal/mol were plotted by starting and stopping indices in the contig. Four plots were created for every contig, one plotted against length of the miRNA sequence, one plotted against the free energy value (in kcal/mol), one plotted against GC content and one plotted against the miRNA number (Plotline function). The output is a .png file for each contig which contain four plots. The maximum of the X axis in each plot is set to the length of each contig in order to accurately show the distribution of the candidate miRNAs in relation to the whole contig, and in order to get this information a .fasta file containing the full sequence of each contig was imported (seqLength function). The GC content of each miRNA was calculated by the proportion of Gs and Cs in the miRNA sequence (CalculateGC).

Results

The output plots show several interesting trends in the candidate miRNAs. Contig 2426 (figure 2) shows a cluster of candidate miRNAs around 33,000. Interestingly, this cluster has an especially high GC content (0.75 – 0.8) compared to the rest of the miRNAs, and to this section of the genome as a whole (which has a GC content of 0.67). This same cluster has similar lengths

and hairpin free energies, which implies that in the GC rich region of the contig, the region that overlaps between all the miRNAs in the cluster forms a stable stem loop structure. A similar trend is seen in contig 62250 (figure 4), which has two very distinct candidate miRNA clusters and several smaller, more dispersed clusters. The miRNAs in the cluster around 15,000 have a GC content above 0.7, but vary greatly in length and free energy. Contig 3109 (figure 3) shows four distinct clusters of miRNAs. This trend is especially apparent on the miRNA number graph, which shows verticle clusters in those regions. These clusters have overlapping regions, but vary widely in length and GC content. Within clusters, their hairpin structure energies are relatively similar. These candidate clusters imply that there is likely a very stable hairpin structure in the overlapping region of each of these clusters. This may imply that theses hairpin structures are more likely to be pre-miRNAs. The previous three contigs all contained many candidate miRNAs, but many of the contigs contained few if any candidate miRNAs. Contig 62250 (figure 5) contains few miRNAs, but shows a distinct clustering of the candidates that were identified within it. MicroRNAs are commonly located in clusters throughout the genome, as more than one miRNA may be expressed from the same primary miRNA transcript (pri-miRNA), and the clustering observed in this contig follows this pattern.

Conclusion

I have compiled a library of candidate microRNA sequences in the C. *mitchellii* genome and analyzed these sequences based on the secondary hairpin structures of their pre-miRNA, then created plots that can be used to analyze this data based on several key parameters. From this data alone, we cannot putatively identify any sequences, but rather provide a starting point for future miRNA identification. Hairpin structure alone is not enough to assign any sort of identification to these sequences, and limiting our results by free energy of these structures may actually exclude possible pre-miRNAs that meet other criteria. Additionally, this data is based on an analysis of only 0.3% of the C. mitchellii genome, so running this program on more of the contigs could reveal important trends in candidate miRNA distribution. Furthermore, a location based approach to analyzing this dictionary of candidates would narrow results and identify clusters. In the future, once the *C. mitchellii* genome is annotated, our list of candidate miRNAs can be further refined to those that fall outside the protein coding region.

68

Start Position : 5241
End Position : 5366
Sequence Size : 126 nucleotides
Minimum Free Energy : -60.5 kcal/mol

Hairpin
((((((....(((.((.((((((((((..((((..(.((((((((((.((.(..((((((........))))))..).)).)))))))))).).)))).....)))))))).)).)))).))))))

Sequence
GUUCUAUUAUGUGAUGCAGGCCUGGGCCUCCGAUGCCUAGAAUGGACUUCCAGAAGUGGCAUCAUG
CACUUCCAGUGGGCCAUUCUAGACCUGGGAGGCCUCCUCCAGGCCUCCAAUACCUAGAAC

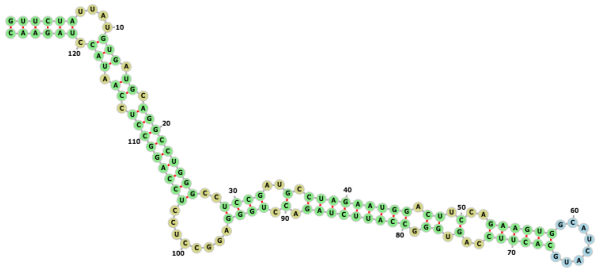

Figure 1. Output file from the miRNA finder web program miRNAFold.

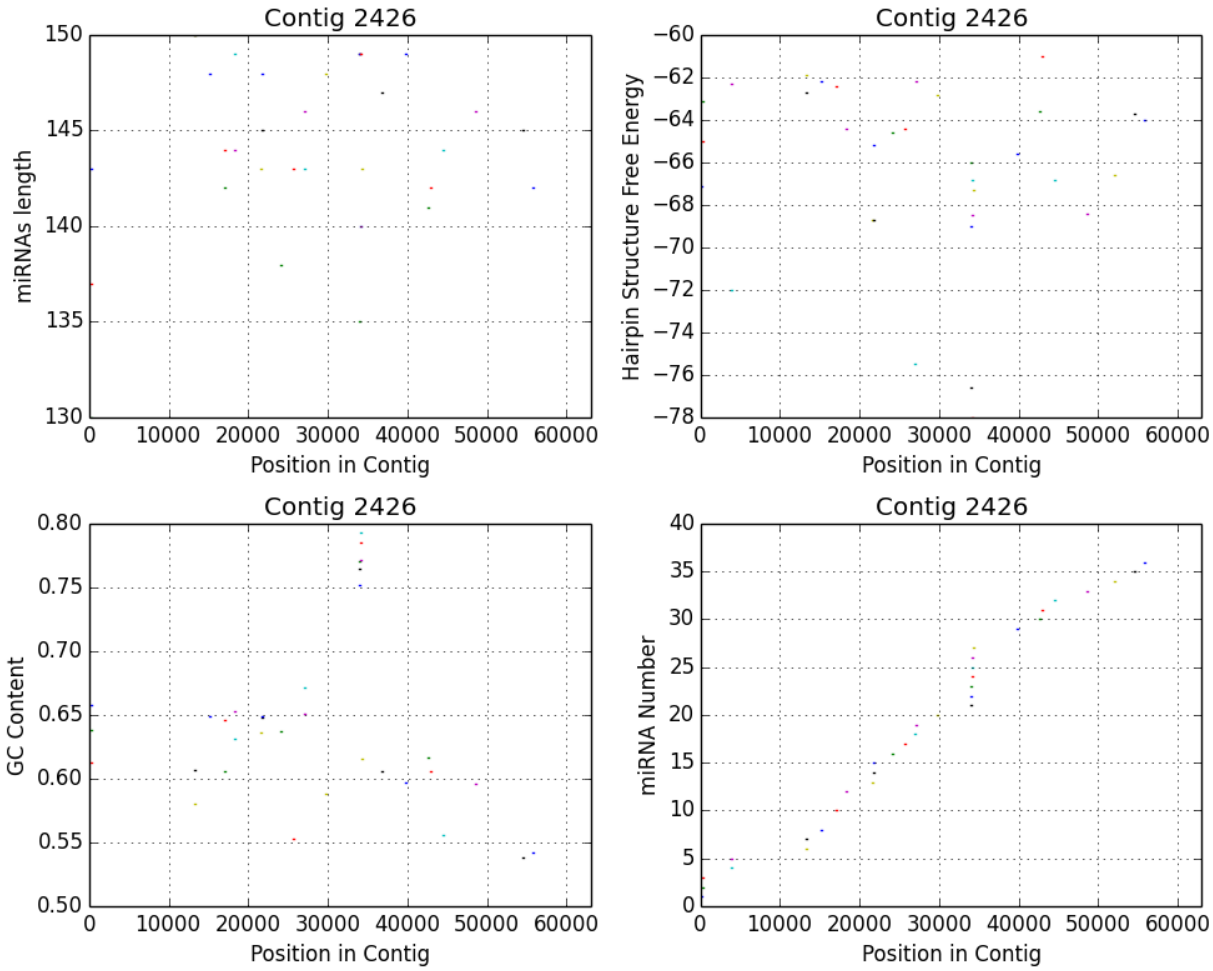

Figure 2. Processed candidate miRNA data created using miRNAPlotMaker from miRNAFold output for contig 2426 from the *C. mitchellii* genome. Hairpin structure free energy is presented in kcal/mol.

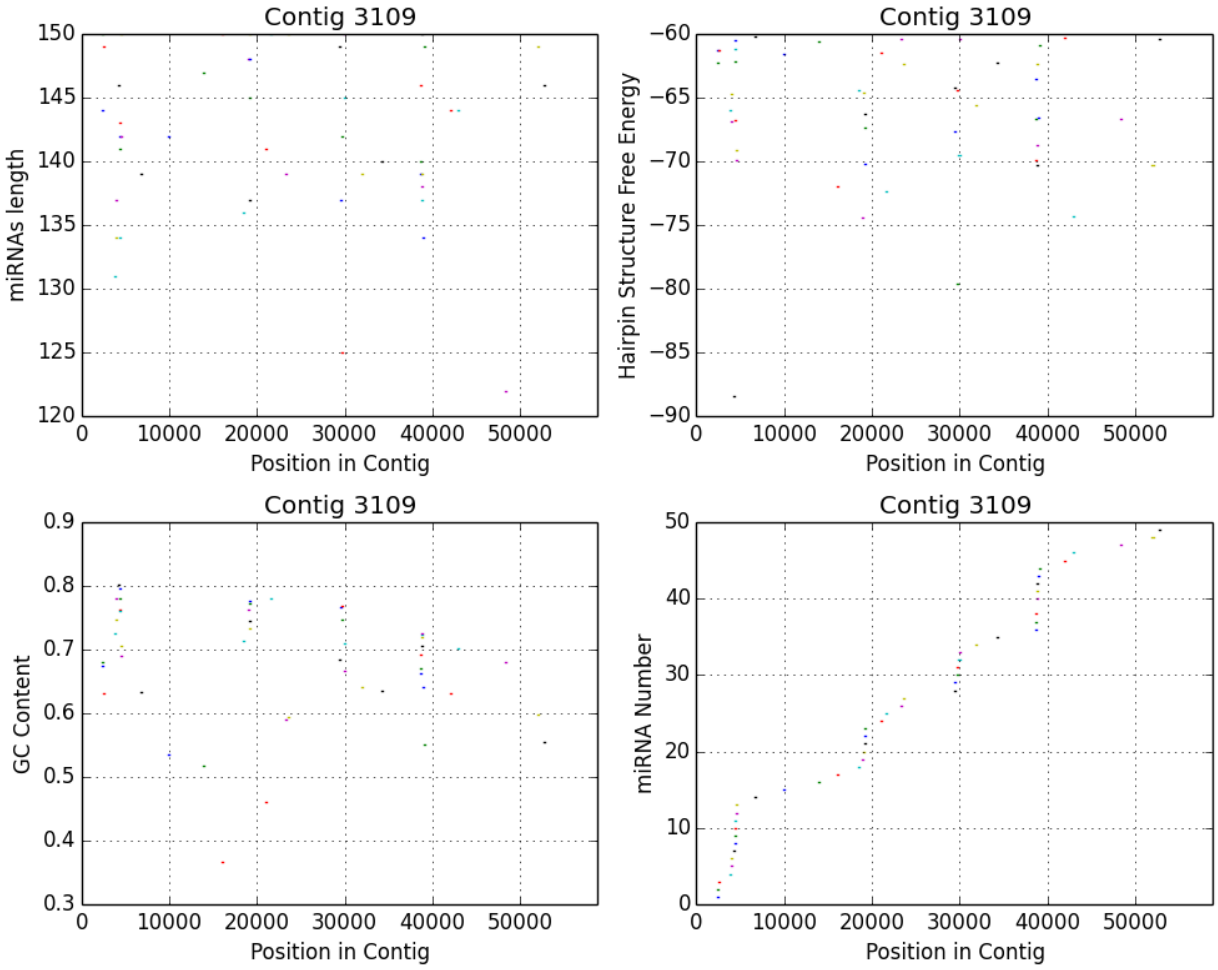

Figure 3. Processed candidate miRNA data created using miRNAPlotMaker from miRNAFold output for contig 3109 from the *C. mitchellii* genome. Hairpin structure free energy is presented in kcal/mol.

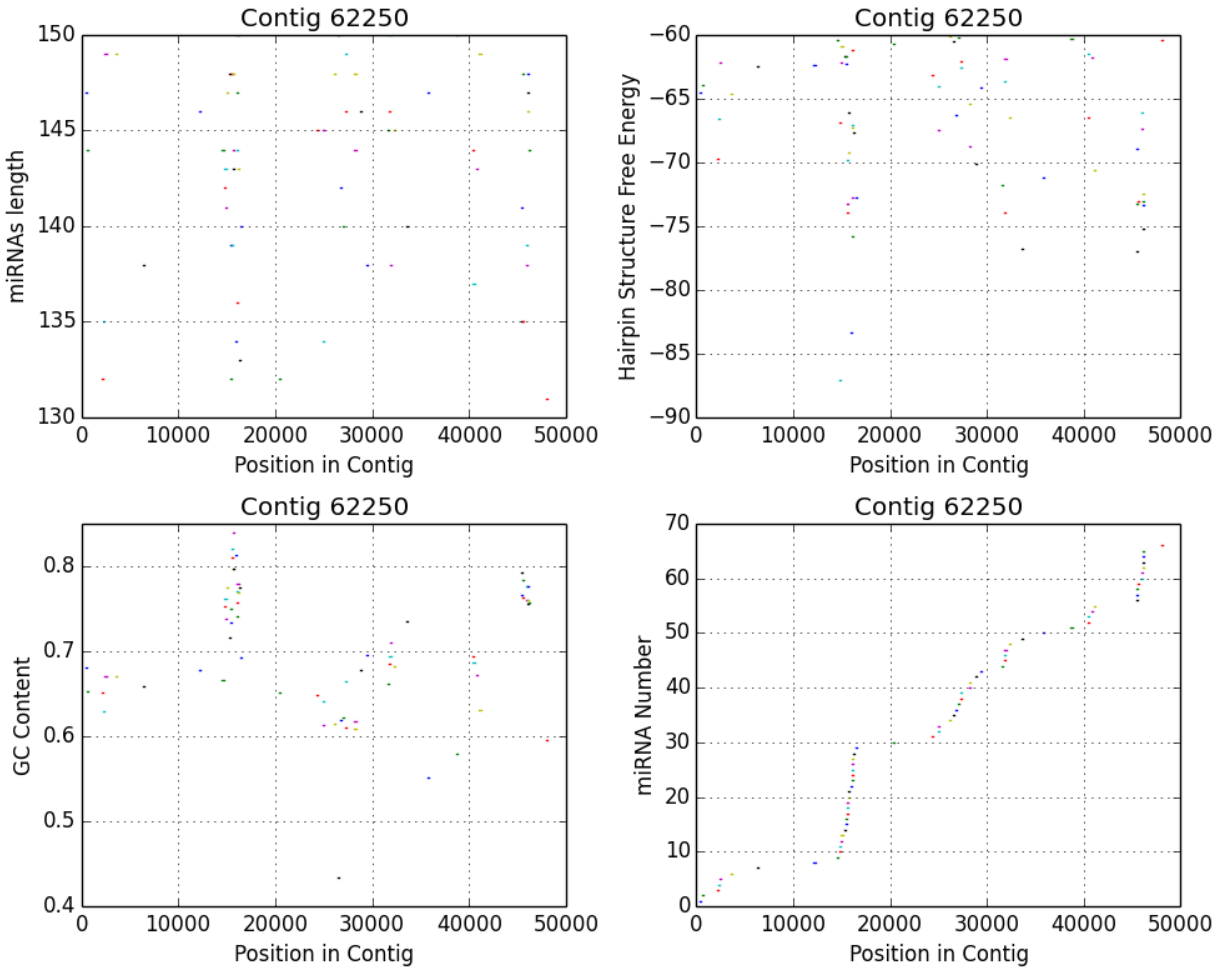

Figure 4. Processed candidate miRNA data created using miRNAPlotMaker from miRNAFold output for contig 62250 from the *C. mitchellii* genome. Hairpin structure free energy is presented in kcal/mol.

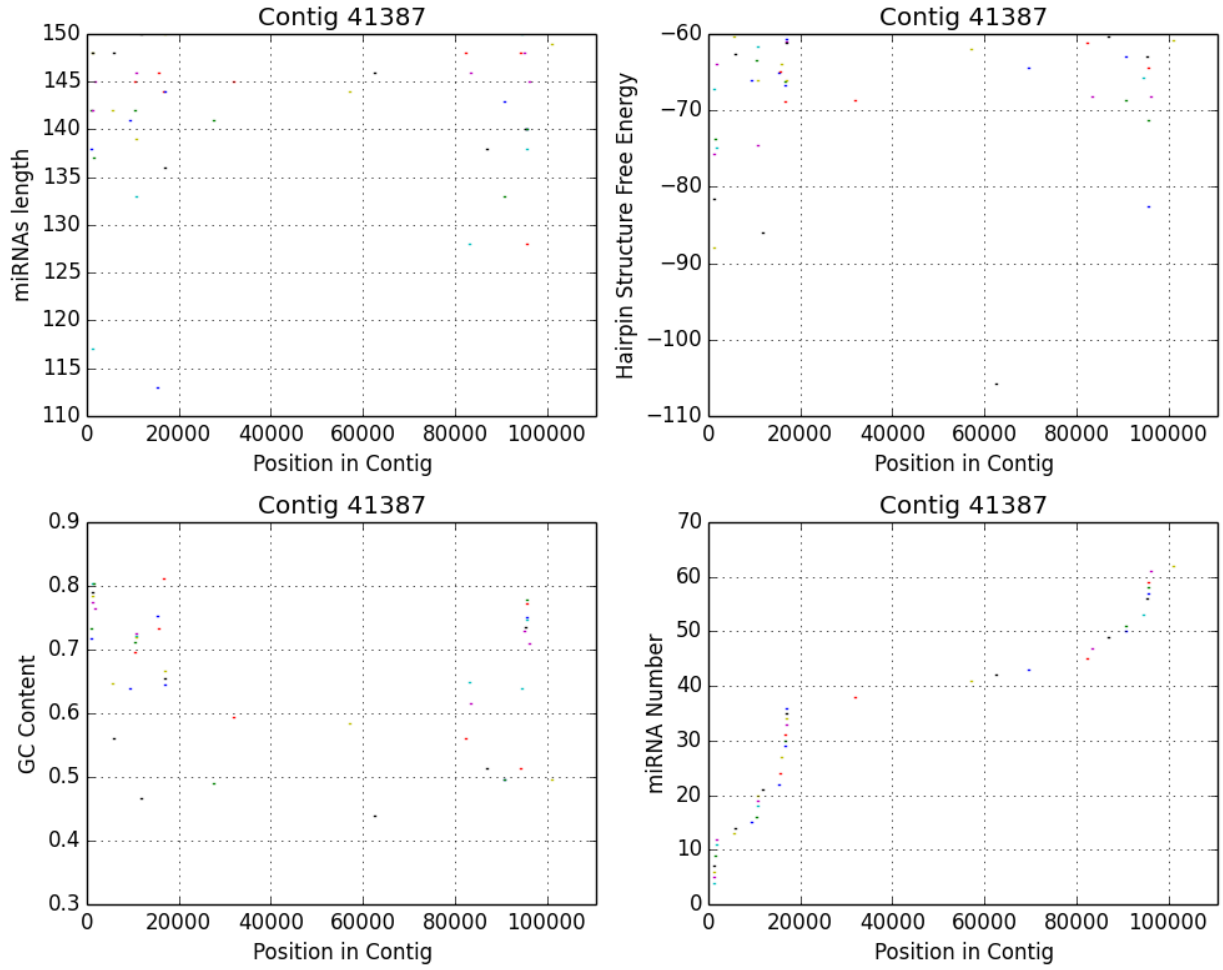

Figure 5. Processed candidate miRNA data created using miRNAPlotMaker from miRNAFold output for contig 41387 from the *C. mitchellii* genome. Hairpin structure free energy is presented in kcal/mol.