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

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

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.

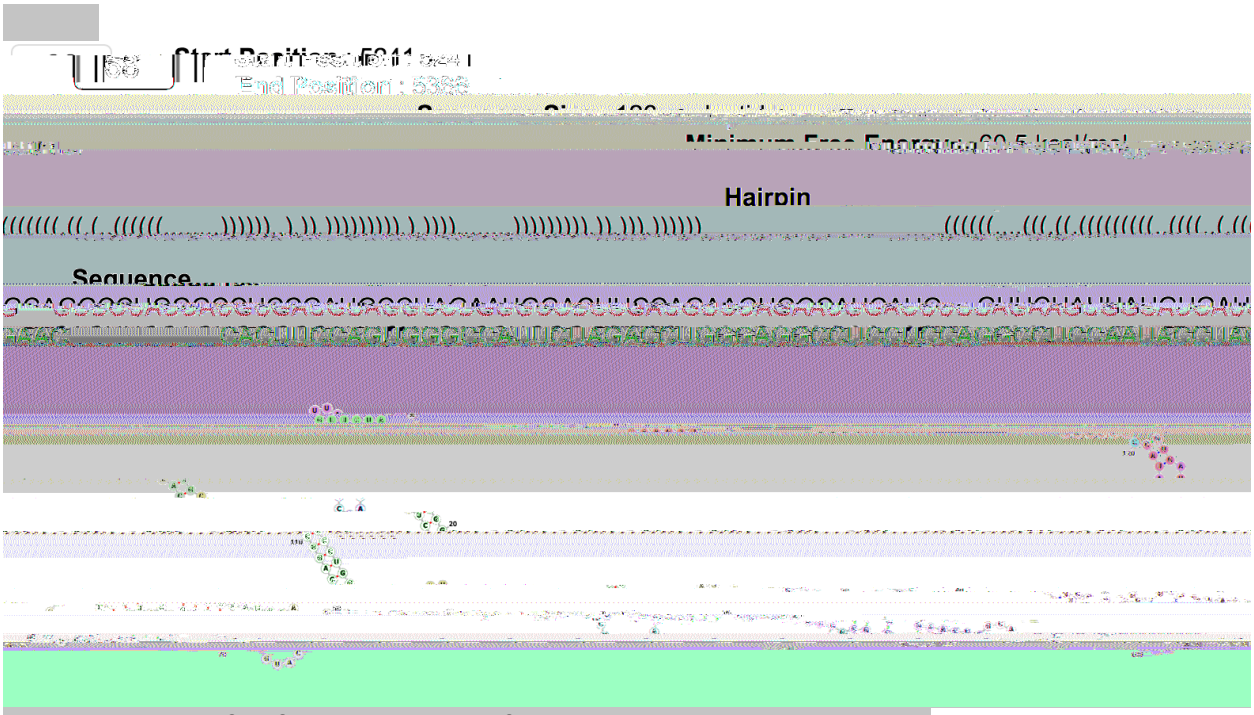


Figure 1. Output file from the miRNA finder web program miRNAFold.

!

"#$%& '!()!*&+, '-- ' .!, /0 .#. /1'!2#345!. /1/!,&' /1' .!%-#0$!2#345*6+17 /8'&!9&+2!2#345"+6.!
+%1 : %1!9+&!, +01#$!(; (<!9&+2!1='! !$'0+2 ')!>/#&:#0!-1&%,1%&'!9&' '!'0'&$?!#-!:&'-'01' .!
#0!8, /6@2+6)!

!

!

!

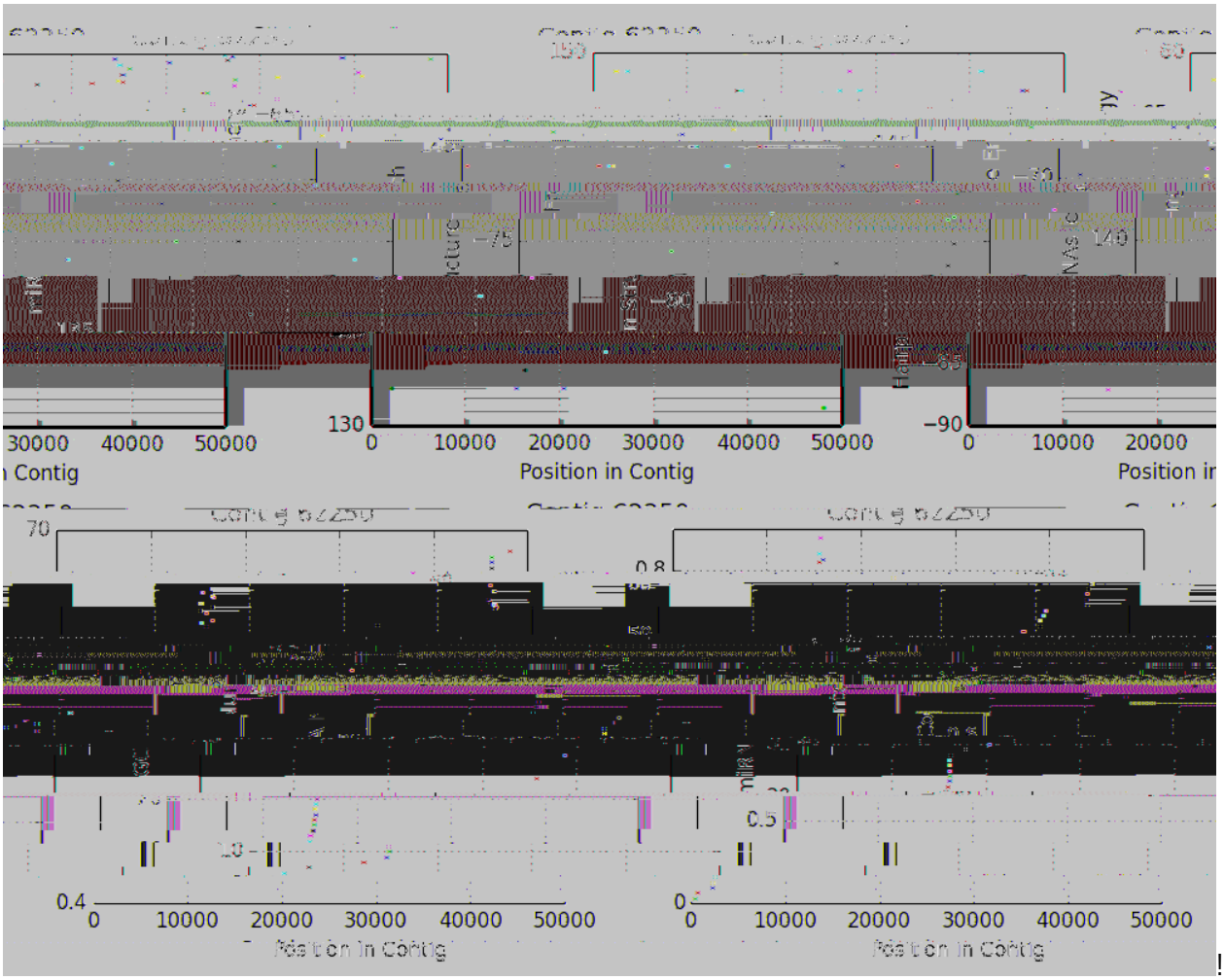
!

!

!

"#$%&'!A)!*&+,'--'.!,/0.#./1'!2#345!./1/!,&'/1'.!%-#0$!2#345*6+17/8'&!9&+2!2#345"+6.!

+%1:%1!9+&!,+01#$!ABCD!9&+2!1='! !$'0+2')!>/#&:#0!-1&%,1%&'!9&''!'0'&$?!#-!:&'-'01'.!



!

"#$%&'!;)!*&+,'--'.!,/0.#./1'!2#345!./1/!,&'/1'.!%-#0$!2#345*6+17/8'&!9&+2!2#345"+6.!
+%1:%1!9+&!,+01#$!<((EC!9&+2!1='! !$'0+2')!>/#&:#0!-1&%,1%&'!9&''!'0'&$?!#-!
:&'-'01'.!#0!8,/6@2+6)!

!

!

!

!

!

!

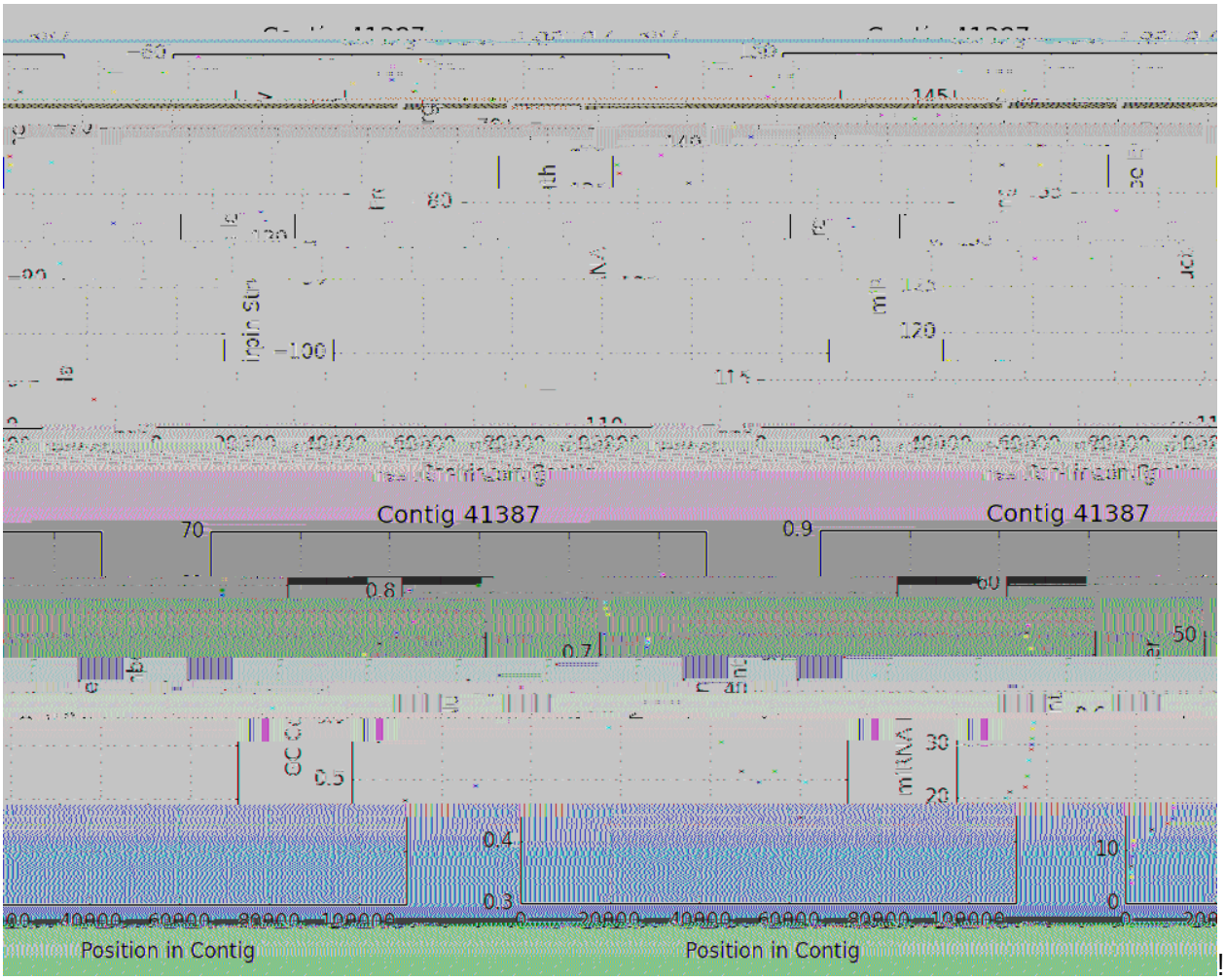
!

!

!

!

!



!

"#$%&'!E)!*&+,'--' .!,/0.#./1'!2#345!./1/!,&'/1' .!%-#0$!2#345*6+17/8'&!9&+2!2#345"+6.!
+%1:%1!9+&!,+01#$!;BAFG!9&+2!1='! !$'0+2')!>/#&:#0!-1&%,1%&'!9&''!'0'&$?!#-!
:&'-'01' .!#0!8,/6@2+6)!

!

!

!

!

!