Positional Preference of Rho-Independent Transcriptional Terminators in E. Coli

Annie Vo

Introduction

Gene expression can be regulated at the transcriptional level through the activities of terminators. Two mechanisms have been established from two distinct types of transcriptional terminators: rho-dependent terminators and rho-independent terminators. Both initiate termination by disrupting the DNA-RNA hybrid complex [2]. Rho-dependent terminators involve the presence of the Rho factor, a protein that binds to a downstream sequence and causes the DNA-RNA complex to dissociate. Rho-independent terminators, also called intrinsic terminators, are sequences that fold into stem-loop structures that also disrupt the complex to terminate transcription and are of interest here. Intrinsic terminators are typically characterized by having a GC-rich stem followed by a T-tract and are energetically favorable to assume the hairpin structure [8].

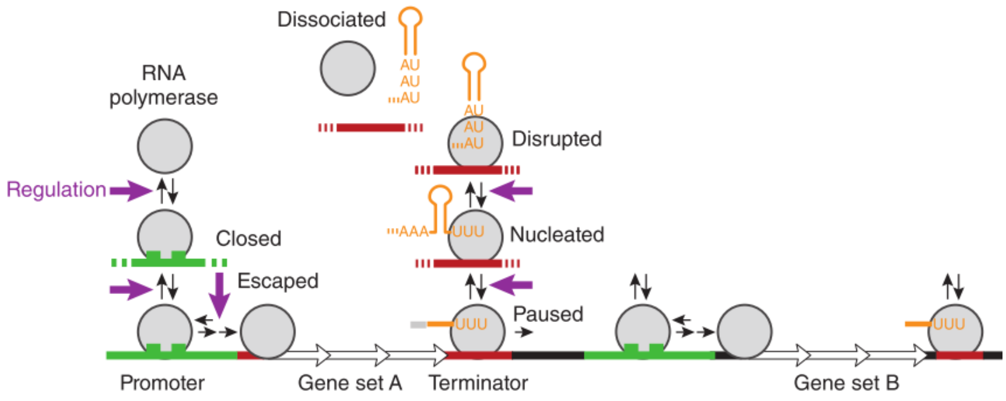


Figure 1 Rho-Independent Terminator. The DNA-RNA complex pauses upon reaching the U-tract, allowing time for the hairpin to form. Completion of the hairpin formation causes a disruption and the complex dissociates, ending transcription. (Adapted from reference 8)

A number of algorithms have been developed to locate terminators from bacterial genomes. Recently, Chen et al. (2013) found 317 natural terminators within the E. coli genome. Theoretically, it would be expected that these terminators lie close to the end of the gene or operon that they regulate in order to prevent needless transcription of RNA and thus to be more energy efficient. However, there are no studies so far that have examined the potential position bias of transcriptional terminators. Thus, this study will examine the distance of intrinsic terminators to determine whether or not terminators exhibit positional preference relative to the end of a gene. A focus will be placed on the highly annotated E. coli genome to help determine why or why not a positional preference may exist.

Only a very specific subset of terminators was examined for this study. The subset consisted of terminators only between parallel genes or converging genes. Also, only terminators on the forward strand. A list of possible terminators was generated using a published algorithm, ARNold. The calculated positions also given with the terminator sequences was then used to determine the distance from the end of a gene en E. coli. To do this efficiently, BioBIKE/PhAnToMe, a web-based programming platform containing annotated genomes was used to calculate the distances. Outliers were viewed by looking right into the genome using other unique functions in BioBIKE/PhAnToMe as well.

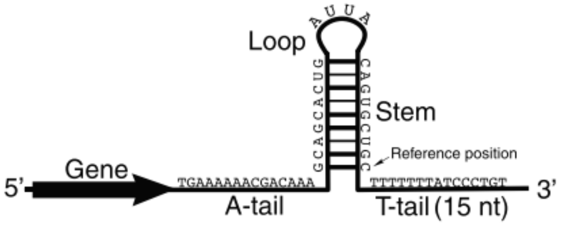
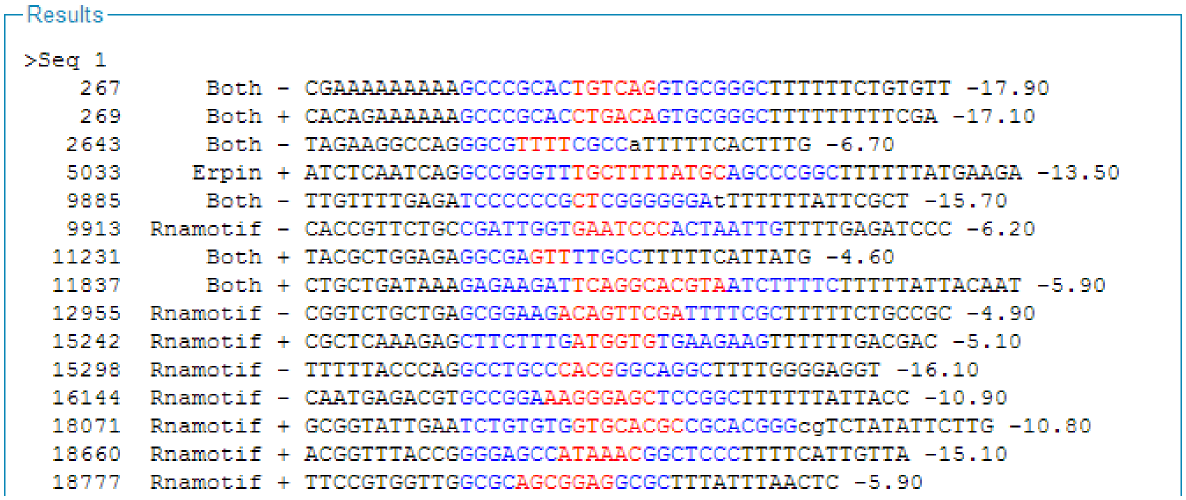


Figure 2 Terminator Distance. The distance from the end of the gene, the length of A-tail denoted, was measured in the study. (Adapted from reference 5)

Methods

Finding terminator sequences

An online web tool called ARNold was used to find possible terminators in E. coli. ARNold utilizes two published algorithms to quickly scan the genome for possible terminators and prints the sequence along with a variety of other results. The two published algorithms used by ARNold are Erpin and RNAMotif [9]. ARNold also provided the 11 nucleotides upstream of the terminator. These upstream sequences were removed when searching for the terminators later.



```
>Seq 1
    267      Both - CGAAAAAAAAAGCCCGCACTGTCAGGTGCGGGCTTTTTTCTGTGTT -17.90
    269      Both + CACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGA -17.10
   2643      Both - TAGAAGGCCAGGGCGTTTTCGCCaTTTTTCACTTTG -6.70
   5033     Erpin + ATCTCAATCAGGCCGGGTTTGCTTTTATGCAGCCCGGCTTTTTTATGAAGA -13.50
   9885      Both - TTGTTTTGAGATCCCCCCGCTCGGGGGGAtTTTTTTATTCGCT -15.70
   9913  Rnamotif - CACCGTTCTGCCGATTGGTGAATCCCACTAATTGTTTTGAGATCCC -6.20
  11231      Both + TACGCTGGAGAGGCGAGTTTTGCCTTTTTCATTATG -4.60
  11837      Both + CTGCTGATAAAGAGAAGATTCAGGCACGTAATCTTTTCTTTTTATTACAAT -5.90
  12955  Rnamotif - CGGTCTGCTGAGCGGAAGACAGTTCGATTTTCGCTTTTTCTGCCGC -4.90
  15242  Rnamotif + CGCTCAAAGAGCTTCTTTGATGGTGTGAAGAAGTTTTTTGACGAC -5.10
  15298  Rnamotif - TTTTTACCCAGGCCTGCCCACGGGCAGGCTTTTGGGGAGGT -16.10
  16144  Rnamotif - CAATGAGACGTGCCGGAAAGGGAGCTCCGGCTTTTTTATTACC -10.90
  18071  Rnamotif + GCGGTATTGAATCTGTGTGGTGCACGCCGCACGGGcgTCTATATTCTTG -10.80
  18660  Rnamotif + ACGGTTTACCGGGGAGCCATAAACGGCTCCCTTTTCATTGTTA -15.10
  18777  Rnamotif + TTCCGTGGTTGGCGCAGCGGAGGCGCTTTATTTAACTC -5.90
```

Figure 3 Sample of ARNold output. Given a genome, ARNold prints the 1) position, 2) algorithm that found the terminator, 3) strand it is located on, 4) terminator sequence, and 5) free energy.

Erpin has a number of built-in profiles that score sequences differently. Profiles can be chosen to search for specific types of sequences. The profiles were designed to be built-in to prevent researchers from having to write specific descriptors; instead, combinations of profiles can be used to make a descriptor to search for specific sequences. For example, when searching for a terminator, sequences are scored using one helix profile and one single-strand profile. A resulting log-odds-score, or a lod-score, is calculated for each of the profiles and the sum of both becomes the overall score for the sequences [4]. High scoring sequences are then returned by the algorithm.

RNAMotif is a highly flexible algorithm that can search for sequences of a wide variety of characteristics. Descriptors are first needed to determine what type of sequence to search for. Once given these descriptors, it searches the genome by each nucleotide to look for matches. Once it finds matches, it can move on to score the sequence in two ways: by thermodynamic stability or by sequence complexity. The thermodynamic scoring accounts for a variety of factors, such as by the free energy to form base pairs and by the free energy of neighboring bases being covalently bonded to one another. The sequence complexity scoring accounts for matches from the descriptors that are similar to highly repeated sequences. Such sequences were typically scored as having low complexity, providing the researcher with a means of determining whether or not the found sequence was a significant match to their descriptor [7].

In ARNold, RNAMotif was given specific descriptors to search for terminator sites. There were 5 descriptors used that were designed by Lesnik et al. (2001). The first noted that the first base of the stem-loop structure could not be an A. The second noted that there must be at least four GC/CG or four GT/TG base pairs in the stem. The remaining descriptors were in regards to the T-tract; there had to be at least 3 Ts closest to the stem, no more than one G, and no more than two adjacent non-T nucleotides within 5 nucleotides. There also could not be more than four purines or 4 cytosines, and there had to be at least four T in the middle and end region of the T-tract [6]. The free energy of the possible terminators was determined by combining the base-stacking and base-pairing free energy. Of the resulting terminators from ARNold, only the terminators found on the forward strand were observed.

Locating terminators

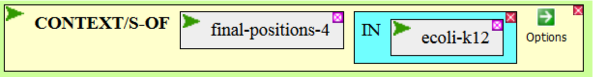
The positions of the possible terminators from ARNold were then used to locate the terminators on the forward strand. Because the positions given in ARNold account for the 11 nucleotides upstream of the terminators, the positions were modified to find the position where the actual terminator started instead. From here, the positions were imported into BioBIKE/PhAnToMe to find the distance of the terminator from the end of a gene.

The function CONTEXT-OF in BioBIKE can return the neighboring genes if given a coordinate. Of the various pieces of information that this function returns, one relates the position to the directionality of the neighboring genes. For example, a terminator can be found between two parallel genes, adjacent genes that are transcribed in the same direction (i.e. 5’ to

3’). Other possibilities are: between converging genes (transcribed in opposite directions towards the other), diverging genes (transcribed in opposite directions away from the other), and inside of a single gene. Because that the terminator does not lie after either gene of a diverging set, these were not considered in this study. Terminators also found in the middle of a gene were also not considered, unless there were overlapping genes. In other words, only terminators between parallel or converging genes were examined in this study.

CONTEXT-OF also provided the distance of the terminator from either gene. Finally, the appropriate distance was chosen depending on the directionality of the genes to obtain the distances of the terminators from the end of the genes.

A)



CONTEXT/S-OF final-positions-4 IN ecoli-k12 Options

B)

```
               3934878
         |--84--|--42--|
>>b3751>>-------*------->>b3752>>
```

C)

```
               3795852
         |--18--|--14--|
>>b3622>>-------*-------<<b3623<<
```

Figure 4 BioBIKE syntax and output. A) The syntax to call the CONTEXT-OF function, where final-positions-4 was a list of positions from the ARNold output; B) a terminator at position 3934878 between two parallel genes; C) a terminator at position 3795852 between converging genes. The values below the position label give the distance from either gene.

Calculations

After obtaining the distance of the terminators from the end of the genes, a few basic statistical values were found. For the set of terminators between converging genes, the mode and median were found; these were done as opposed to the average to avoid having the data skewed by outliers. The range was found to determine the extent of the outliers. The same values were found for the terminators between parallel genes, and then for the set of all of the genes.

Results

	Terminators	Median	Mode	Range
Converging	199	26	10, 13	2 – 506
Parallel	376	35	11, 33	1 – 3527
Overall	575	33	10, 19	1 – 3527

Table 1 Distance (in nucleotides) of Terminators from Gene End. The median distance of terminators from the end of genes found separately by gene directionality and then as a whole. The most reoccurring distances were also found alongside the ranges.

From ARNold, 3248 possible terminators were found. Of these, only those on the forward strand and between either parallel or converging genes were examined. This reduced the number of observed terminators to 575. Of the observed, there were about 1.5 times more terminators between parallel genes than between converging genes. Overall, there does not appear to be any significant difference between the distances of terminators between converging or parallel genes. Most terminators appear to be 10 nucleotides or 19 nucleotides from the end of the gene, and the median distance of terminators from genes is 33 nucleotides.

Discussion

Transcriptional terminators do appear to have a positional preference relative to the end of a gene. Both the parallel and converging genes were consistent in showing that the most repeated distance from a gene was 10 nucleotides. However, both sets also had another distance that was highly repeated, and it was a much greater distance in the parallel set than the converging set. Similarly, the median distance was also higher in the parallel set. A part of this may be attributed to the fact that there were 177 more terminators between parallel genes than converging genes. This may also be the cause of a much larger range seen in the parallel set. Further studies can be done to determine whether or not these differences can be attributed to the gene directionality or simply the size of the sets. As a whole, it is not understood why this positional preference exists

There were a number of strange cases in this data set. Only terminators read in the positive direction were examined in this study. However, a large number of these terminators were found after genes that were read in the opposite direction. For example, the 1415485 terminator is between two parallel genes that are read in the negative direction. However, this terminator is read in the positive direction; it follows a gene for a phage protein and precedes a gene for exodeoxyribonuclease VIII. Because this is unlikely to be a part of the reason the terminator is in the opposite directionality of the gene, perhaps looking at the free energy of the terminator might shed light on the matter. It is possible that the terminator itself is unlikely to fold and in fact does not end transcription of the phage protein gene. Because there are many other cases like this, a further study as a whole could be done to look at terminators in the opposite directionality of parallel genes.

Because that the distances covered such a wide range in both sets, a closer look was taken at some of the outliers. The first case observed is the terminator at position 4209084 between two parallel genes; the terminator comes 1433 nucleotides after the gene, which encodes for rRNA. Interestingly, this gene overlaps with another gene that codes for an LSU rRNA; both genes start at the same position, and the LSU rRNA gene is much longer. As such, this also means that the 4209084 terminator is in the middle of a different gene. This may contribute to why the terminator is found extremely far downstream of from the shorter, overlapping rRNA gene. A future study looking into terminators that occur in the middle of genes can expand on this; perhaps other terminators excluded from the examined set were also a part of overlapping genes. Another terminator at position 4499663 is between converging genes and is 506 nucleotides after the end of the gene read in the positive direction. Closer examination revealed that the gene it follows encodes a hypothetical protein. Investigation into the identity of the protein may explain why this distance to the terminator is larger than usual.

While this study does show that transcriptional terminators exhibit a degree of positional preference, there are still many studies that can be done to expand on the topic. The reason as to why terminators may exhibit a positional bias is still not understood. The functions of the genes with terminations a specific distance could be examined to see if there is a relationship between the function of the gene and the distance of the terminator. Terminators that are a shorter distance from the end of the gene may follow genes that help maintain homeostasis in the organism and are therefore needed more frequently. Similarly, because E. coli is an organism that utilizes operons, a study looking for a relation between operons and the position of terminators could be done. A terminator for an operon might explain why it is placed further away from the end. This study also looked at a very specific subset of terminators; there were issues calculating the position of terminators on the other strand in this study, and thus the terminators on the negative strand were not examined. However, including them in the examined set could provide more insight into the question of why a positional bias is observed. Additionally, all terminators in this study were specifically found after the end of a gene. It did not account for terminators that might have minor overlaps with the end of the gene, such as a case where the terminator is located 2 nucleotides before the end of the gene. On a broader scale, bidirectional terminators were not looked into at all. Studies into any of the aforementioned could lead to a better understanding of how genes are regulated.

References

1. Chen, Ying-Ja et al. "Characterization of 582 Natural and Synthetic Terminators and Quantification of Their Design Constraints." *Nature Methods* 10.7 (2013): 659–64. Web. 20 Mar. 2014.

2. d'Aubenton Carafa, Y, E Brody, and C Thermes. "Prediction of Rho-Independent Escherichia Coli Transcription Terminators. A Statistical Analysis of Their RNA Stem-Loop Structures." *Journal of Molecular Biology* 216.4 (1990): 835–58.

3. Ermolaeva, M D et al. "Prediction of Transcription Terminators in Bacterial Genomes." *Journal of Molecular Biology* 301.1 (2000): 27–33. Web. 4 Apr. 2014.

4. Gautheret, D, and a Lambert. "Direct RNA Motif Definition and Identification from Multiple Sequence Alignments Using Secondary Structure Profiles." *Journal of Molecular Biology* 313.5 (2001): 1003–11. Web. 8 May 2014.

5. Kingsford, Carleton L, Kunmi Ayanbule, and Steven L Salzberg. "Rapid, Accurate, Computational Discovery of Rho-Independent Transcription Terminators Illuminates Their Relationship to DNA Uptake." *Genome Biology* 8.2 (2007): R22. Web. 17 Apr. 2014.

6. Lesnik, E a et al. "Prediction of Rho-Independent Transcriptional Terminators in Escherichia Coli." *Nucleic Acids Research* 29.17 (2001): 3583–94.

7. Macke, T J et al. "RNAMotif, an RNA Secondary Structure Definition and Search Algorithm." *Nucleic Acids Research* 29.22 (2001): 4724–35.

8. Mooney, Rachel Anne, and Robert Landick. "Building a Better Stop Sign: Understanding the Signals That Terminate Transcription." *Nature Methods* 10.7 (2013): 618–619. Web. 21 Mar. 2014.

9. Naville, Magali et al. "ARNold: A Web Tool for the Prediction of Rho-Independent Transcription Terminators." *RNA Biology* 8.1 (2011): 11–13. Web. 8 Apr. 2014.

All Distances:

```
[1, 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8,

8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,

10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12,

12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14,

14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 17, 17, 17,

17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19,

19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,

21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23,

23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26,

26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28, 28, 28, 28,

28, 28, 28, 28, 29, 30, 30, 30, 30, 30, 30, 31, 31, 31, 31, 31, 32, 32, 32, 32, 32, 32, 32, 32,

32, 32, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 33, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34,

34, 35, 35, 35, 35, 35, 35, 35, 35, 35, 36, 36, 36, 36, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,

37, 37, 37, 38, 39, 40, 40, 40, 41, 41, 42, 42, 42, 42, 42, 42, 42, 42, 43, 43, 43, 43, 43, 43,

43, 44, 45, 45, 45, 45, 45, 45, 45, 47, 47, 47, 47, 47, 47, 47, 48, 48, 48, 48, 49, 50, 50, 50,

50, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 53, 53, 53, 53, 55, 55, 55, 57, 57, 58, 59, 59, 59,

59, 59, 59, 59, 59, 61, 62, 62, 62, 63, 63, 64, 64, 65, 67, 67, 68, 68, 69, 69, 71, 71, 71, 72,

73, 74, 74, 75, 76, 77, 77, 77, 78, 78, 79, 80, 83, 84, 84, 84, 84, 85, 85, 87, 87, 90, 90, 95,

96, 96, 97, 99, 101, 102, 102, 103, 105, 106, 107, 107, 108, 109, 110, 111, 113, 116, 116, 121,

122, 124, 124, 124, 125, 128, 130, 131, 132, 133, 134, 134, 136, 138, 138, 139, 139, 139, 141,

141, 141, 148, 148, 152, 152, 153, 156, 156, 156, 158, 158, 165, 166, 166, 167, 168, 168, 170,

170, 171, 171, 173, 173, 173, 175, 176, 176, 177, 177, 183, 187, 187, 187, 187, 187, 187, 187,
```

188, 191, 192, 193, 194, 194, 196, 201, 204, 205, 206, 218, 220, 231, 231, 231, 231, 232, 239, 240, 243, 262, 277, 283, 296, 296, 298, 298, 310, 311, 317, 317, 337, 341, 353, 361, 361, 378, 378, 383, 390, 423, 425, 428, 440, 451, 472, 506, 506, 530, 674, 703, 1433, 1433, 3527]

Counts for Each Distance (distance: counts):

```
Counter({10: 19, 18: 17, 23: 15, 13: 14, 27: 14, 7: 13, 37: 13, 11: 12, 14: 12, 21: 12, 34: 12,
12: 11, 28: 11, 33: 11, 32: 10, 9: 9, 15: 9, 22: 9, 24: 9, 26: 9, 35: 9, 8: 8, 17: 8, 19: 8, 25:
8, 42: 8, 59: 8, 20: 7, 43: 7, 45: 7, 47: 7, 187: 7, 30: 6, 51: 6, 16: 5, 31: 5, 6: 4, 36: 4,
48: 4, 50: 4, 52: 4, 53: 4, 84: 4, 231: 4, 4: 3, 5: 3, 40: 3, 55: 3, 62: 3, 71: 3, 77: 3, 124:
3, 139: 3, 141: 3, 156: 3, 173: 3, 1: 2, 2: 2, 3: 2, 41: 2, 57: 2, 63: 2, 64: 2, 67: 2, 68: 2,
69: 2, 74: 2, 78: 2, 85: 2, 87: 2, 90: 2, 96: 2, 102: 2, 107: 2, 116: 2, 134: 2, 138: 2, 148: 2,
152: 2, 158: 2, 166: 2, 168: 2, 170: 2, 171: 2, 176: 2, 177: 2, 194: 2, 296: 2, 298: 2, 317: 2,
361: 2, 378: 2, 1433: 2, 506: 2, 29: 1, 38: 1, 39: 1, 44: 1, 49: 1, 58: 1, 61: 1, 65: 1, 72: 1,
73: 1, 75: 1, 76: 1, 79: 1, 80: 1, 83: 1, 95: 1, 97: 1, 99: 1, 101: 1, 103: 1, 105: 1, 106: 1,
108: 1, 109: 1, 110: 1, 111: 1, 113: 1, 121: 1, 122: 1, 703: 1, 125: 1, 128: 1, 130: 1, 131: 1,
132: 1, 133: 1, 136: 1, 153: 1, 674: 1, 165: 1, 167: 1, 175: 1, 183: 1, 188: 1, 191: 1, 192: 1,
193: 1, 196: 1, 201: 1, 204: 1, 205: 1, 206: 1, 218: 1, 220: 1, 232: 1, 239: 1, 240: 1, 243: 1,
530: 1, 262: 1, 277: 1, 283: 1, 310: 1, 311: 1, 337: 1, 341: 1, 353: 1, 383: 1, 390: 1, 423: 1,
425: 1, 428: 1, 440: 1, 451: 1, 3527: 1, 472: 1})
```

Converging Terminators: 199

Median - 26.0

Mode - [(10, 13)]

Range - 2 to 506

Parallel Terminators: 376

Median - 35.0

Mode - [(33, 11)]

Range - 1 to 3527

Total forward terminators: 575

Median - 33.0

Mode - [(10, 19)]

Range - 1 to 3527