Problem: Alignment of tRNA gene models
Input: Intersection of genes predicted by Aragorn and tRNAscan-SE

Method:
step1: Removing the variable arms. I used the sequences reported by TSE since they are all reported up to position 73. using the secondary structure reported by TSE for each sequence, I counted number of arms. if there were four arms, I removed the third arm.
Exp:

```
BayalaiB08-376_Baya_001_4_S
">>>>>>>..>>>..........<<<.>>>>>.......<<<<<.>>>...<<<..>>>>>.......<<<<<<<<<<<<<."
"GTCACCATACCCAAGTGGTtACGGGGACTGACTAGAAATCAGTTGCGATCTCGCGCGCAGGTTCGAATCCTGCTGGTGACG"
after removing variable arm:
"GTCACCATACCCAAGTGGTtACGGGGACTGACTAGAAATCAGTTGCGCAGGTTCGAATCCTGCTGGTGACG"
">>>>>>>..>>>..........<<<.>>>>>.......<<<<<...>>>>>.......<<<<<<<<<<<<<."
```

We had no sequences with less than three arms in our intersection set with 3580 sequences.

Step2: Removing introns and insertions, by removing all the lower case letters. (Based on TSE manuscript, in genes found by tse nucleotides matching the "consensus" tRNA model used in Cove analysis appear in upper case, while introns and other nucleotides in non-conserved positions are printed in lower-case letters.)
Exp:

```
CfasciculataCfCl_CfaC1_25_76_Y
"CCTTCTGTAGCTCAATTGGTAGAGCATGTGACTGTAGagttagcagatATCACAGGGtCGCTGGTTCGATTCCGGCCGGAAGGA"
">>>>>>>..>>>>.......<<<<.>>>>>...................<<<<<.....>>>>>.......<<<<<<<<<<<<<."

after removing the introns and insertions:
"CCTTCTGTAGCTCAATTGGTAGAGCATGTGACTGTAGATCACAGGGCGCTGGTTCGATTCCGGCCGGAAGGA"
```

Step3: Running covea with eukaryotic model.

Step4: Editing the covea's output by removing:
1. Sites that have more than 99% gap
2. Sequences with more than 8 gaps
3. Sequences which contain letter n or N
   (the secondary structure reported by covea is also updated along with every step)

Step5: Mapping the sites to sprinzl coordinates. This step is done manually!
Here is the final structure of the aligned sequences of Intersection set with length 72:

```
#=CS    >>>>>>>..>>>>.......<<<<.>>>>>.......<<<<<....>>>>>.......<<<<<<<<<<<<<.
```

By comparing the structure to sprinzl coordinates. I was able to match all the sites to sprinzl coordinates.
The Exceptions were in:
- **Loop D:** We have 8 bases in this loop. I assumed positions 17,17A (which are not always present) from sprinzl were missing in our structure. Because position 17 in our alignment was relatively invariant (3017 sequences out of 3580 sequences had letter G at this position) which fits position 18 from sprinzl. So, one of the positions 20A or 20B should be present.
- **Anticodon-arm:** Positions 26 and 44 were not base paired and the arm has only 4 base pairs.
- **Position 47** from sprinzl (which is not always present) is also not present in our structure.

I integrated the script for every step into one script which you can find here.

Step6: Validating the alignment. I haven't decided how to validate the results yet!
Step7: Alignment of ~700 tRNA models predicted by ARA only. I need to use ARA secondary structure which I haven't done yet!