INTRODUCTION

Tranfer-messenger RNA (tmRNA) is one among the vast universe of RNAs whose versatile roles have been studied to shred lights on important transcriptional-related processes. Most famously known from such is its reparatory function of stalled ribosomes at the end of faulty mRNA, which lacks of stop codons and can, lethal for the cell, by binding with Small protein B (SmpB) protein, forming a complex that slips on mRNA and ending transcription when its own stop codon is reached. As suggested by the name, one can expect tmRNA also posesses both tRNA and mRNA functions besides rescuing stalled ribosomes. Reportedly, these include inducing transpeptidation when it binds to a A-site on ribosome, or controlling of mRNA and protein quality during transcription. tmRNA is exclusively found in nearly all eubacteria and some organelles like chloroplasts. The study of tmRNA and highly conservative regions across various bacteria strains is as important as fascinating for many future prospects.

The degree of difference between two or many sequences, be it DNA, RNA or amino acids, can be effectively measured by different types of distance calculating methods. For comparing strains of the same length, it is simple and sufficient to use Hamming distance, while in the opposite case, edit distance should be calculated using more complicated dynamic programming methods. In our case, the Hamming distance reveals how many RNA nucleotides are different between two sequences.

The base pair distance, defined by the number of base pairs present in one structure and absent in all others, is also target of discussion. By examining Hamming distance and base pair distance for Bacillus tmRNA sequence.

MATERIALS AND METHODS

The FASTA file containing aligned tmRNA sequences from Bacillus is procured from <https://rth.dk/resources/rnp/tmRDB/rna/tmrna.html>. The first sequence is the pairing mask that map the base pairing positions in the tmRNA using different keys for each resulting helix loops. In other sequences, the corresponding positions, if are in uppercase, show the bases involving in a specific loop. The tmRNA sequences in this alignment also contain gaps, indicated by ‘-‘, and unknown bases, indicated by ’n’, which are to be filtered before pairwise distance calculation.

For parsing the FASTA file, SeqIO is used and each sequence with its ID is stored. For the extraction of base pairs, all positions of a loop type and corresponding base are stored in a Python dict whose keys are loop types. The positions where the pairing start and the length of pairing are used to trace out the sequences in the pair.

The base pair distance is calculated by finding the number of base pair present in one sequence but not the other, and vice versa. Another measurement of distance is Hamming distance, where the pair of full length sequences is aligned. The distance in positions is incremented where there is a mismatch at non-gap and non-unknown bases positions. Subsequently, the histograms of base pair distance and Hamming distance are generated to investigate the similarity in tmRNAs extracted from a wide range of Bacillus bacteria.

RESULTS

1. Extraction of all base pairs

All base pairs are extracted based on the pairing mask and are stored for each individual sequence according to their type of loop. They can be accessed using the function extract\_basepair(sequence\_id).

1. Calculation of base pair distance

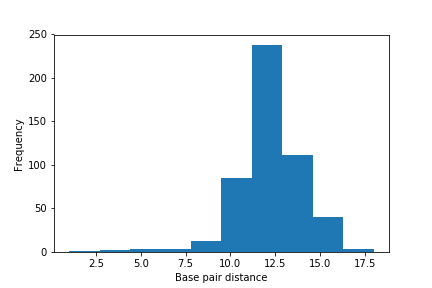


Figure 1. Histogram of base pair distance

The base pair distance reflects how different the loop-forming base pair in 2 sequences are. Most pairs are different by 10-14 base pairing sequences. More extreme values are less observed.

1. Calculation of Hamming distance

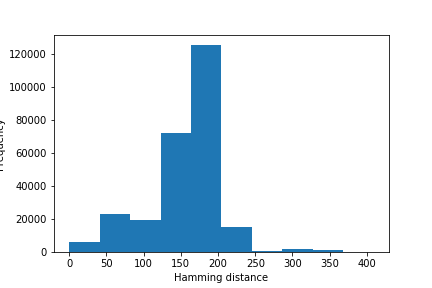
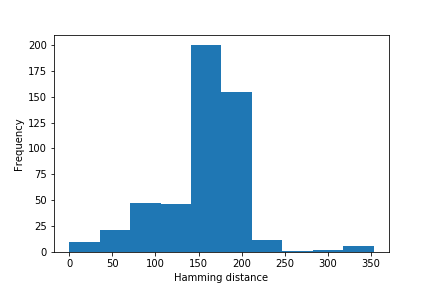


Figure 2. Histogram of Hamming distance for 500 and all pairs of sequences

For 500 randomly selected pair of sequences, the frequency of pairwise Hamming distance is illustrated in Figure 2. Most pairs are different at approximately 150-200 bases and while the length of each alignment is 1341, this means the error has the coverage of about 11-15%. This is understandable as the sequences are derived from different Bacillus strains and thus, there are a number of highly conservative regions involve, while they can vary each other greatly in non-conservative regions.

Another important observation is that the number of pairs whose Hamming distance smaller than 150 is larger than that whose distance larger than 200. Because the sequences all come from Bacillus bacteria, there is higher chance that they are more similar to each other, rather than different. Again, the conservation of tmRNA structure among the same is demonstrated.

The frequency of Hamming distance for all possible pairs is also plotted. In exception for the distribution of error between 150-200 bases, the same trend is observed for the regions smaller than 150 and larger than 200.

DISCUSSION

The target of our study is tmRNA sequences extracted from different Bacillus strains. Two methods for measurement of distance between two sequences: the base pair distance and Hamming distance. Here we discuss about strength and weaknesses of the method.

The Hamming distance is a widely used method and in this practical, it proves to be helpful as discussed above: We can see the Bacillus strains are more likely to have high similarity to each other, indicated by the skewness of the histogram to the left where the distance is smaller. Very few pairs of sequences are significantly different (distance of >300 base pairs). This shows the conservativeness of essential parts in tmRNA in the evolutionary pathways of different bacteria strains. Since we have the prepared alignment of equal-length segments, the Hamming distance can be applied. Nonetheless, if the sequences are not yet aligned and of various length, other methods like calculating edit distance and score functions might be more effective in handling gaps and mismatches, giving a more accurate distance.

The base pair distance, although do reflect the difference, do not illustrate the aspects showed by Hamming distance. We can study further about the variation of helix loop structure formed by different base pairs, but a lot more investigation should be done in addition. In the other hand, base pair distance is a simple way to examine the pairwise difference.

CONCLUSION

tmRNA of Bacillus bacteria are studied by base pair distance and Hamming distance. While Hamming distance is more powerful than base pair distance, it should be taken into consideration that both methods have fines and flaws. The study is meaningful for further analysis of tmRNA structures and functions, as well as the conservativeness of the sequence in evolution.