Chapter 5

Base substitutions in tRNA gene secondary structure motifs

5.1 Introduction

In DNA, there are four different transition (*ti*) substitutions in which a purine (or a pyrimidine) base is replaced by another purine (or a pyrimidine) base (R→R; Y→Y). Similarly, there are eight different transversion (*tv*) substitutions in which a purine (or a pyrimidine) base is replaced by a pyrimidine (or a purine) base (R→Y; Y→R) [44] Figure 5-1. Under conditions where all the twelve substitutions are occurring in equal proportions, the ti/tv ratio would be 0.50. But, the reported ratio in different genomes is usually ~ 2.00 or more, which suggests that a transition is $\sim$ four times more frequent than a transversion in DNA [39][123][187][204]. This bias towards transition has been known since comparison between homologous DNA sequences [56][238]. The purine:purine and pyrimidine:pyrimidine pairing distorts the geometry of the double helix for which transversion frequency is lower than that of transition [199][217]. The lower transversion frequency can also be due to several other factors. One such factor is cytosine deamination in DNA during replication and transcription causing one of the most common transitions, C→T (G→A) [105][219]. In coding regions, transversions at the third and first codon

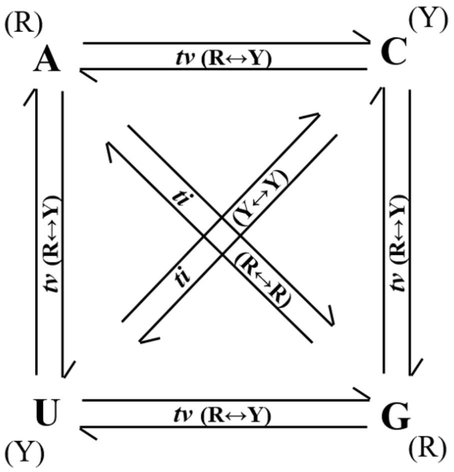
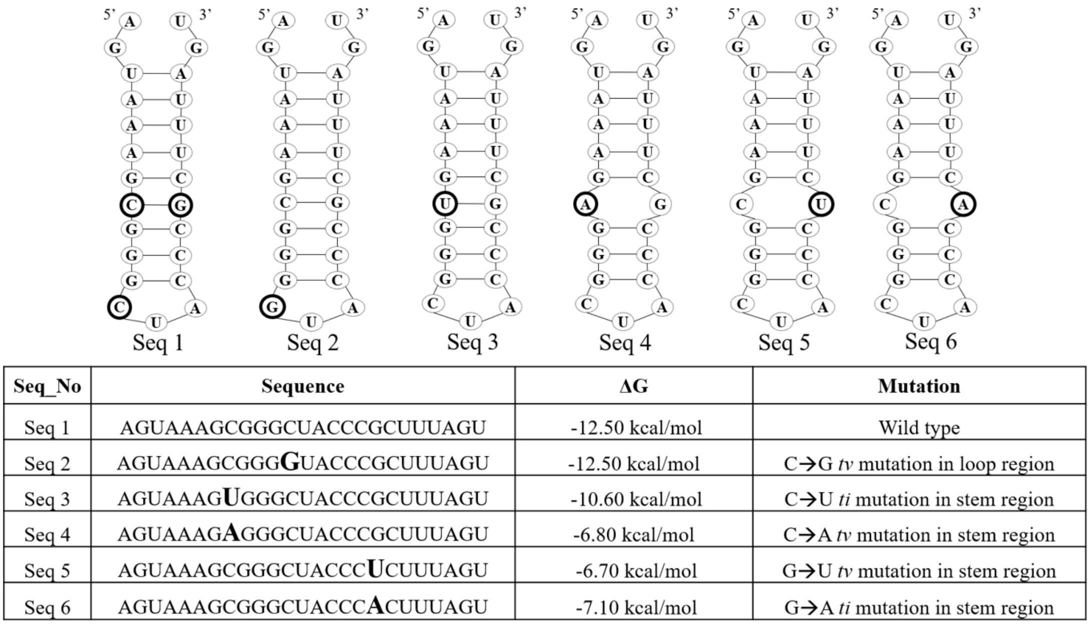


Figure 5-1: Different substitution mutations in the genome

Figure presents twelve possible directional base substitutions in a sequence. In theory, out of the four bases A, C, G, and U any one base can be replaced by the other three bases resulting into twelve base substitutions. Out of these twelve substitutions, the four are called transitions (*ti*) in which a purine (R: A/G) (or a pyrimidine (Y:C/U)) is replaced by another purine (R) (or a pyrimidine (Y)); eight different substitutions called transversion (*tv*) in which a purine (R) (or a pyrimidine (Y)) is replaced by a pyrimidine (Y) (or a purine (R))

positions are more non-synonymous than transition [1][226].

Though it is well known that *ti* is more frequent than *tv* in genomes, all factors influencing their occurrence in genomes remains to be explored further. Transition and transversion were observed with similar frequencies at non-methylated cytosine sites in grasshopper pseudo-genes [92]. Selective constraints imposed by secondary structure can account for the relative enrichment of *ti* in tRNA and rRNA genes [95]. In case of tRNA and rRNA genes, there is selection for G+C enrichment in the stem region unlike in the loop region, in thermophilic bacteria [51][70][228]. It is known that RNA secondary structure plays an important role in gene expression and regulation [144]. There can be different impacts of transition and transversions in the encoded RNA that forms double helix as explained with the help of schematic scenarios as described in Figure 5-2. For example, a transition such as G→A results in A:C pairing while the A→G substitution results in G:U pairing in the encoded transcript. A transversion such as G→U (R→Y) results in U:C (Y:Y) pairing while U→G (Y→R) substitution results in G:A (R:R) pairing in the transcript. The different base pairing due to *ti* and *tv* substitutions can cause different magnitudes of instabilities in the secondary structure of a transcript [173]. The selection pressure on secondary structure acts in favor of transition and against the transversion, which is reflected in changes in free energy level (Figure 5-2). Therefore, it can be assumed that *tv* and *ti* mutations are fixed/selected according to evolutionary constraints associated with maintenance of tRNA secondary structure.



Seq_No	Sequence	ΔG	Mutation
Seq 1	AGUAAAGCGGGCUACCCGCUUUAGU	-12.50 kcal/mol	Wild type
Seq 2	AGUAAAGCGGG**G**UACCCGCUUUAGU	-12.50 kcal/mol	C→G *tv* mutation in loop region
Seq 3	AGUAAAG**U**GGGCUACCCGCUUUAGU	-10.60 kcal/mol	C→U *ti* mutation in stem region
Seq 4	AGUAAAG**A**GGGCUACCCGCUUUAGU	-6.80 kcal/mol	C→A *tv* mutation in stem region
Seq 5	AGUAAAGCGGGCUACCC**U**CUUUAGU	-6.70 kcal/mol	G→U *tv* mutation in stem region
Seq 6	AGUAAAGCGGGCUACCC**A**CUUUAGU	-7.10 kcal/mol	G→A *ti* mutation in stem region

Figure 5-2: Effect of substitutions on secondary structure in a hypothetical RNA sequence

Figure presents secondary structure and minimum free energy for six hypothetical sequences (wild type (Seq 1) and five others with base substitutions (Seq 2-Seq 6)). Base substitutions are marked in bold. RNAfold webserver [64][116][128][129] available at http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi was used for estimating secondary structure and minimum free energy (ΔG) of these sequences. Seq 2 having a base substitution in the loop region and the Seq 1 have equal minimum free energy suggesting the maximum stability. Seq 3 and Seq 6 have the transitions in stem region. Secondary structures of Seq 3 and Seq 6 have lesser free energy than Seq 4 and Seq 5, having transversions. Predicted folded structure as well as their free energy support the hypothesis that transversion is more deleterious for the RNA secondary structure than the transition. The C→T (or A→G) transition is more favorable than T→C (or G→A) transition.

Earlier studies on tRNA genes regarding *ti* and *tv* have mainly been carried out by comparing genes across the species [71][80][90][183]. The main finding is that compensatory transition or transversion substitutions are more frequent than single site independent substitution in stem regions of these genes. Further, *ti* to *tv* ratio is observed to be higher in stem region in comparison to loop region as compensatory *tv* is more rare than compensatory *ti.* However, a systematic study of *ti* and *tv* in tRNA by comparing gene sequences within a species have not been reported in literature till date. Considering single substitution being more frequent than double substitutions, the possible occurrence of non-compensatory substitutions in stem region cannot be avoided if the study is carried out within a species. In addition, as G:U base pairing often being allowed in tRNA stem, non-compensatory substitution from an amino base to keto base might be preferred over the reverse substitution, which is yet to be explored. Availability of genome sequences from a large number of strains belonging to a species has opened up avenue to address the above queries. tRNA structures are relatively easy to survey for finding out the stem and the loop regions in different organisms. Therefore, in this study, we carried out a comparative analysis of *ti* and *tv* substitutions in tRNA genes using large whole genome datasets of five bacterial species: *Escherichia coli (Ec), Klebsiella pneumoniae (Kp), Salmonella enterica (Se), Staphylococcus aureus (Sa) and Streptococcus pneumoniae (Sp).* The frequency of *ti* was found to be higher than *tv* in the stem regions than in the loop region of tRNA genes. Further the transitions from amino to keto bases were found to be more frequent than the reverse transitions in the stem regions. These observations indicate that secondary structure in tRNA influences differentially the transition and transversion frequencies in organism.

5.2 Materials and Methods

5.2.1 Extracting intergenic regions, tRNA genes and segregating loop and stem regions

In this study, we have considered five bacteria, *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Staphylococcus aureus* (*Sa*), *Salmonella enterica* (*Se*) and

Streptococcus pneumoniae (*Sp*) for which large number of genome sequences were available in public databases. In total we have done a detailed computational analysis of the tRNA genes of the alignment of 157 *Ec* strains [215], 208 *Kp* strains [76], 132 *Sa* strains [165], 366 *Se* strains [215], and 264 *Sp* strains [31] for finding out base substitutions. Python scripts were written to extract alignments of the tRNA gene sequences from the alignment of DNA sequences by using coordinate information of the annotated tRNA genes. Intergenic regions (IRs) which are the stretches of DNA sequences located between genes are also extracted considering coordinates of the coding regions (protein coding genes, tRNA genes and rRNA genes). Genomic compositional details of the tRNA genes and IRs analyzed are given in Table 5.1.

We extracted the predicted secondary structure of the tRNA genes using tRNAscan-SE On-line webserver [117] available at http://lowelab.ucsc.edu/tRNAscan-SE/. This web database (GtRNAdb) is rich in genomic tRNA information. It confirms the clover leaf shaped tRNA secondary structure, gene size, anti-codon position and anti-codon for a given query tRNA gene sequence. The extracted tRNA sequences from the alignments of genomes but not confirmed in the GtRNAdb, were not analyzed in this study. The extracted genes of *Ec*, *Kp*, *Sa*, *Se* and *Sp* are 89, 86, 61, 88 and 59 respectively, but confirmed number of genes from GtRNAdb of the organisms are 87, 85, 61, 87 and 58 respectively. Using this webserver, we obtained predicted secondary structure of the tRNAs in terms of dot bracket notations given in the Appendix A.4.5 and the list of tRNA genes of five bacteria is presented in Appendix A.4.6. For further analysis, we considered these predicted secondary structures and classified nucleotides into two groups (i) paired and (ii) unpaired. Paired and unpaired nucleotides are considered largely to be from stem and loop regions, respectively. In general, the paired nucleotides or the stem region are the bases occurring in the acceptor stem, D-arm, anticodon-arm, variable region and T-arm. The unpaired nucleotides of the above four arms and the variable region are considered as the

Table 5.1: ti/tv ratio in tRNA genes and Intergenic Regions (IRs)

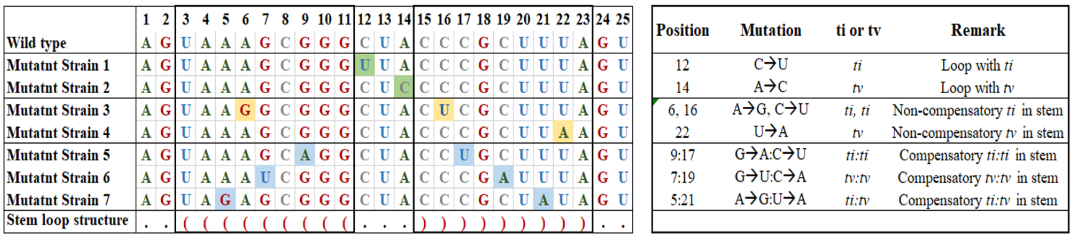
Genome (G+C%) /Size	Genomic Regions	Size$^{\#}$	*ti*	*tv*	*ti*+*tv*	*ti* / Size	*tv* / Size	(*ti*+*tv*)/ Size	ti/tv
Ec (50.79/4641652)	IRs	526035	29649	18549	48198	0.056	0.035	0.092	1.598
	tRNA (87)*	6827	46	19	65	0.007	0.003	0.010	2.421
	tRNA (stem)	3734	22	9	31	0.006	0.002	0.008	2.444
	tRNA (stem non-compensatory)	3734	20	5	25	0.005	0.001	0.007	4.000
	tRNA (stem compensatory)	3734	2	4	6	0.001	0.001	0.002	0.500
	tRNA (loop)	3093	24	10	34	0.008	0.003	0.011	2.400
Kp (57.68/5248520)	IRs	579525	30785	20351	51136	0.053	0.035	0.088	1.513
	tRNA (85)*	6668	24	11	35	0.004	0.002	0.005	2.182
	tRNA (stem)	3648	8	0	8	0.002	0.000	0.002	NA$^{\$}$
	tRNA (stem non-compensatory)	3648	6	0	6	0.002	0.000	0.002	NA$^{\$}$
	tRNA (stem compensatory)	3648	2	0	2	0.001	0.000	0.001	NA$^{\$}$
	tRNA (loop)	3020	16	11	27	0.005	0.004	0.009	1.455
Sa (32.83/2832299)	IRs	592925	25296	20715	46011	0.043	0.035	0.078	1.221
	tRNA (61)*	4705	52	14	66	0.011	0.003	0.014	3.714
	tRNA (stem)	2618	29	9	38	0.011	0.003	0.015	3.222
	tRNA (stem non-compensatory)	2618	21	7	28	0.008	0.003	0.011	3.000
	tRNA (stem compensatory)	2618	8	2	10	0.003	0.001	0.004	4.000
	tRNA (loop)	2087	23	5	28	0.011	0.002	0.013	4.600
Se (52.19/4879400)	IRs	559086	59367	26949	86316	0.106	0.048	0.154	2.203
	tRNA (87)*	6820	76	30	106	0.011	0.004	0.016	2.533
	tRNA (stem)	3728	36	10	46	0.010	0.003	0.012	3.600
	tRNA (stem non-compensatory)	3728	28	4	32	0.008	0.001	0.009	7.000
	tRNA (stem compensatory)	3728	8	6	14	0.002	0.002	0.004	1.333
	tRNA (loop)	3092	40	20	60	0.013	0.006	0.019	2.000
Sp (39.49/2221315)	IRs	282088	19514	9785	29299	0.069	0.035	0.104	1.994
	tRNA (58)*	4376	30	14	44	0.007	0.003	0.010	2.143
	tRNA (stem)	2460	22	1	23	0.009	0.000	0.009	22.000
	tRNA (stem non-compensatory)	2460	16	1	17	0.007	0.000	0.007	16.000
	tRNA (stem compensatory)	2460	6	0	6	0.002	0.000	0.002	NA$^{\$}$
	tRNA (loop)	1916	8	13	21	0.004	0.007	0.011	0.615

*Number in the parentheses denotes number of tRNA genes of the bacteria analyzed in this study
# Size of the genomic region analyzed in this study
$Value not available
Frequency of polymorphism between IRs and tRNA is significantly different (p-value < 0.01)
Frequency of polymorphism between IRs and tRNA stem is significantly different (p-value < 0.01)
Frequency of polymorphism between IRs and tRNA loop is significantly different (p-value < 0.01)
Frequency of polymorphism between tRNA stem and tRNA loop is not significantly different (p-value > 0.01)
Frequency of polymorphism between tRNA non-compensatory stem and tRNA loop is not significantly different (p-value > 0.01)

bases in the loop regions (Appendix A.4.1).

5.2.2 Segregating compensatory and non-compensatory substitutions in stem regions

For each tRNA gene, substitution positions were mapped to the secondary structure and segregated into loop and stem regions. Further, the substitutions in stem region were marked as compensatory or non-compensatory depending on whether there exist a pair of substitutions or only one substitution in a paired position in the stem region respectively (Figure 5-3).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Wild type	A	G	U	A	A	A	G	C	G	G	G	C	U	A	C	C	C	G	C	U	U	U	A	G	U
Mutatnt Strain 1	A	G	U	A	A	A	G	C	G	G	G	U	U	A	C	C	C	G	C	U	U	U	A	G	U
Mutatnt Strain 2	A	G	U	A	A	A	G	C	G	G	G	C	U	C	C	C	C	G	C	U	U	U	A	G	U
Mutatnt Strain 3	A	G	U	A	A	G	G	C	G	G	G	C	U	A	C	U	C	G	C	U	U	U	A	G	U
Mutatnt Strain 4	A	G	U	A	A	A	G	C	G	G	G	C	U	A	C	C	C	G	C	U	U	A	A	G	U
Mutatnt Strain 5	A	G	U	A	A	A	G	C	A	G	G	C	U	A	C	C	U	G	C	U	U	U	A	G	U
Mutatnt Strain 6	A	G	U	A	A	A	U	C	G	G	G	C	U	A	C	C	C	G	A	U	U	U	A	G	U
Mutatnt Strain 7	A	G	U	A	G	A	G	C	G	G	G	C	U	A	C	C	C	G	C	U	A	U	A	G	U
Stem loop structure	.	.	(	(	(	(	(	(	(	(	(	.	.	.	)	)	)	)	)	)	)	)	)	.	.

Position	Mutation	ti or tv	Remark
12	C→U	*ti*	Loop with *ti*
14	A→C	*tv*	Loop with *tv*
6, 16	A→G, C→U	*ti, ti*	Non-compensatory *ti* in stem
22	U→A	*tv*	Non-compensatory *tv* in stem
9:17	G→A:C→U	*ti:ti*	Compensatory *ti:ti* in stem
7:19	G→U:C→A	*tv:tv*	Compensatory *tv:tv* in stem
5:21	A→G:U→A	*ti:tv*	Compensatory *ti:tv* in stem

Figure 5-3: Compensatory or non-compensatory polymorphisms in secondary structure of a hypothetical RNA sequence

Figure presents hypothetical scenarios of base substitutions in loop, compensatory and non-compensatory polymorphisms in stem regions (polymorphisms are depicted in left panel and descriptions are given in right panel). Wild type is the original sequence and there are 7 mutant strains. The secondary structure of the sequence is in dot-bracket notation, where dot represents a base in loop and bracket represents a base in stem. In mutant strain 1 and 2, there are one *ti* and one *tv* polymorphisms at the 12th and the 14th positions in the loop region respectively. In mutant strain 3, there are two non-compensatory *ti* polymorphisms in stem at the 6th and the 16th positions. In mutant strain 4, there is one non-compensatory *tv* polymorphism at the 22nd position. Mutant strain 5, 6 and 7 has compensatory polymorphisms. In mutant strain 5, there are two *ti* polymorphisms at the 9th and the 17th positions, in mutant strain 6, there are two *tv* polymorphisms at the 7th and 19th positions, respectively. In mutant strain 7, there is one *ti* polymorphism in one side and one *tv* polymorphism in the other side of compensatory pair at 5th and 21st positions, respectively.

5.2.3 Finding substitutions from the sequence alignments

Considering the most frequent nucleotide at a position in the alignment of the nucleotide sequences of a tRNA gene, we computed a reference sequence and then used this reference sequence to identify a substitution in each sequence [215]

(Appendix A.2.1). Substitution frequencies were computed by dividing total count of a given substitution by the total number of the nucleotide in which substitution has occurred. For example, suppose the total number of C→U substitution is 2 and total number of C in a tRNA gene sequence is 10, then the normalized frequency would be $2/10 = 0.2$. We further wrote a Python script to classify these substitutions into transitions (ti) and transversions (tv). Observed substitutions in secondary structure of a sample tRNA gene are shown in Appendix A.4.1. We further classified the substitutions in stem regions as compensatory and non-compensatory. For statistical analysis and determining p-value for significance test, Mann Whitney test is used [125].

5.2.4 Visualization of 2-D and 3-D structures of tRNA genes

For 2-D visualization of tRNA secondary structure we have used tRNAscan-SE On-line software [117]. To visualize the tRNA 3-D structures, we have used two web servers. First, we gave the tRNA sequence and the secondary structure in dot bracket notation obtained from tRNAscan-SE On-line software as input to Vfold3D webserver [242][251], to obtain the 3-D structure in pdb format. Next, the 3-D secondary structure of tRNA was visualized from the pdb file using iCn3D web server [230] (Appendix A.4.1).

5.3 Results

5.3.1 Higher transition to transversion ratio in tRNA genes than intergenic regions

By multiple sequence alignment from hundreds of strains of a species, the possible twelve substitutions were found in intergenic regions (IRs) and tRNA genes of five bacterial species such as *Ec*, *Kp*, *Sa*, *Se* and *Sp* (Table 5.1). Substitution frequency in tRNA genes was observed to be ten times lower than that in the IRs. The difference between IRs and tRNA genes is significant (p-value < 0.01) in all the five bacteria. This finding was anticipated because the tRNA genes are transcribed to make functional tRNA structures which carries out the vital translation process inside the cell. Therefore, the low substitution frequency in tRNA genes is most likely due to strong purifying selection on these genes in comparison to IRs. We compared *ti* and *tv* between the two regions. The *ti/tv* values were greater than 1.0 across the five bacteria, which suggested that the *ti* frequency is more than that of *tv* in both the genomic regions. The *ti/tv* values in tRNA genes were greater than that in the IRs (p-value < 0.05) (Table 5.1). There might be two possibilities for the higher *ti/tv* value in tRNAs: either low *tv* or high *ti*. We calculated the relative fold increase in the *ti* and *tv* in the IRs, separately, in comparison to the tRNA genes (Appendix A.4.7). The fold increase of *tv* was more than that of *ti* in the IRs (p-value < 0.05), which suggested that the higher *ti/tv* in tRNA is due to lower *tv* occurrence in tRNA than that in IRs. This observation was in support of the notion that impact of *tv* on tRNA secondary structure is higher than that of *ti* (Figure 5-2).

5.3.2 Higher transition to transversion ratio at the stem regions than the loop regions within tRNA genes

Transfer RNA genes have well defined secondary structures: the double stranded helical stem regions that constitutes $\sim$ 2/3rd part and the single stranded loop regions that constitutes $\sim$ 1/3rd part of a tRNA gene. We analyzed substitutions separately in stem and loop regions. The substitutions in both loop as well as stem regions were observed to be significantly lower than that in the IRs. This difference indicated that both the regions are under strong purifying selection as mentioned above. It is known that the stem and the loop regions are functionally important and make contacts with translation factors. To compare the stem and the loop regions regarding *ti* and *tv*, we separated substitutions in the stem regions as compensatory as well as non-compensatory substitutions (Table 5.1). It was evident that non-compensatory substitutions were more in number than compensatory substitutions (Appendix A.4.2). This is pertinent to note that in previous studies tRNA genes were compared across species for which often the substitutions were observed as compensatory, unlike the observed values here. It may be noted that compensatory substitutions are relatively earlier in evolution in comparison to non-compensatory substitutions considering the higher stability of the former than the latter. We considered the ti/tv values of non-compensatory substitutions in the stem region to compare it further with the substitutions in the loop regions. The ti/tv values in the stem region were observed to be significantly higher than that in the loop regions across the bacteria (p-value < 0.01) (Figure 5-4). This suggested that frequency of *tv* is more in the loop region than the stem region. This observation was in concordance with our hypothesis that *tv* in stem regions are more deleterious than the *ti*.

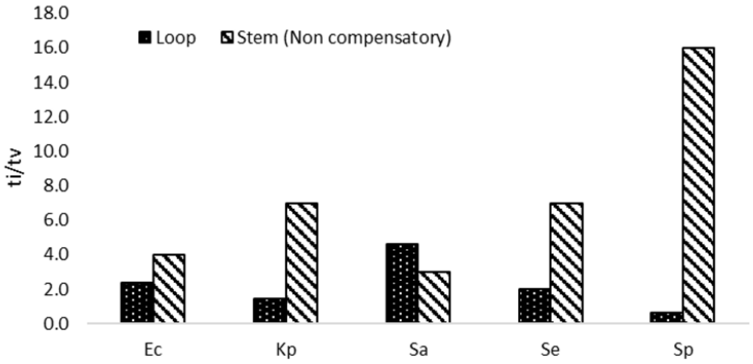


Figure 5-4: Ratio of *ti* to *t*v in loop and non-compensatory stem regions in tRNA of five bacteria

Histogram presenting the ratio of *ti* to *tv* (ti/tv) values in loop and stem regions in tRNA genes. Only non-compensatory substitutions are considered in stem region. The ti/tv values between stem and loop tRNA genes are significantly different (p-value < 0.05). The x-axis presents the five bacteria *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Salmonella enterica* (*Se*), *Staphylococcus aureus* (*Sa*) and *Streptococcus pneumoniae* (*Sp*). To avoid division by 0 error, *ti* and *tv* values are incremented by 1 each in case of *Kp*.

5.3.3 Biased transition substitution towards keto bases in the stem region of tRNA genes

In tRNA stem regions, the G:U pairing is found to be accepted more favorably than A:C pairing. Therefore, non-compensatory substitutions from amino bases (A/C) to keto bases (G/T) that facilitates G:U pairing are likely to be favorable in the stem region. However, the reverse transition such as non-compensatory substitutions from keto bases (G/T) to amino bases (A/C) that facilitates A:C pairing are likely to be less preferred in the stem region. These two pairings have been described in Appendix A.4.3 and the stability of these pairing in terms of interaction energy has been calculated using GAUSSIAN 09 software [48] which are presented in Appendix A.4.8. This encouraged us to compare the stem and the loop regions further in terms of substitutions from amino bases to keto bases and the vice versa. In stem regions, substitutions from amino (A/C) to keto bases (G/T) were significantly more than keto to amino bases (p-value < 0.05) (Figure 5-5). This high amino to keto substitutions were consistently observed among all the five bacteria studied. However, in loop regions, no such significant pattern was observed with regard to substitutions from keto to amino or amino to keto bases (Figure 5-6). This provided additional support that secondary structure influences significantly towards different substitutions in tRNA genes. Further, the comparative results between keto to amino and amino to keto transitions in IRs indicated that, there was no significant difference between these two transition substitutions in IRs (p-value > 0.01) (Appendix A.4.4).

Isoacceptor tRNAs are with different anticodons but charged with the same amino acid by the same amino acyl tRNA synthetase enzyme. Therefore, we explored to compare among the isoacceptor tRNA genes regarding substitution frequencies. In majority of the cases, number of substitutions per tRNA gene was observed as 0 or 1 or 2 (Appendix A.4.9). Therefore, an elaborate comparison among these iso- acceptors tRNA genes regarding substitutions was not possible. However,

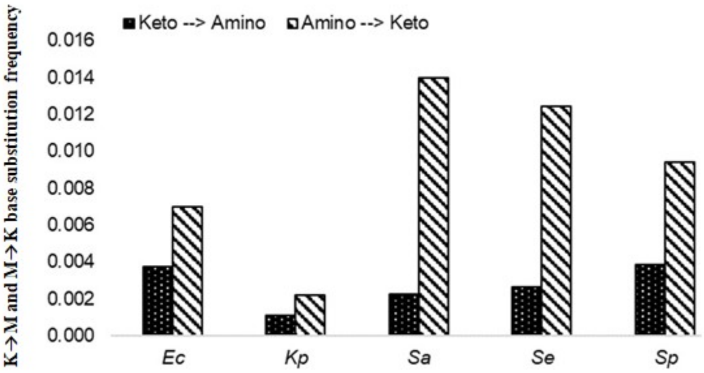


Figure 5-5: Polymorphism frequencies in stem region of tRNA genes among Amino(A/C) → Keto (G/T), Keto (G/T) → Amino (A/C)

Histogram presenting the frequency values of single nucleotide polymorphism in stem regions in tRNA genes. Only non-compensatory substitutions are considered in stem region. The normalized values between Amino → Keto, Keto → Amino polymorphism frequencies in stem regions of tRNA genes are significantly different (p-value < 0.05). The x-axis presents the five bacteria *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Salmonella enterica* (*Se*), *Staphylococcus aureus* (*Sa*) and *Streptococcus pneumoniae* (*Sp*).

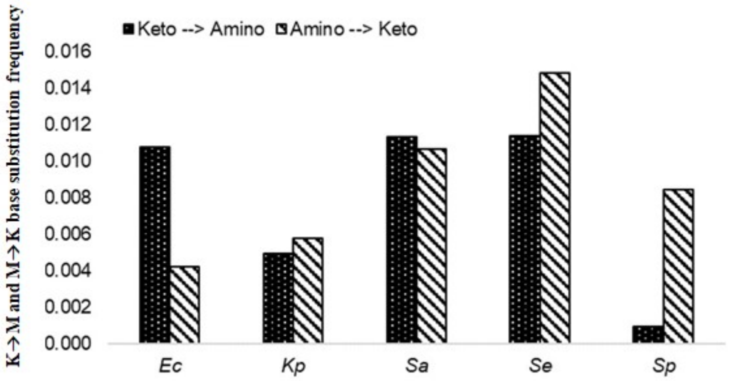


Figure 5-6: Amino(A/C) → Keto (G/T), Keto (G/T) → Amino (A/C) polymorphism frequencies in loop region of tRNA genes

Histogram presenting the frequency values of single nucleotide polymorphism in loop regions in tRNA genes. The normalized values between Amino (A/C) → Keto (G/T), Keto → Amino polymorphism frequencies in loop regions of tRNA genes are not significantly different (p-value > 0.05). Further there is no consistent pattern here across the bacteria, unlike the stem region. The x-axis presents the five bacteria *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Salmonella enterica* (*Se*), *Staphylococcus aureus* (*Sa*) and *Streptococcus pneumoniae* (*Sp*).

there were a few cases where the number of substitutions among isoacceptor tRNA genes have noticeable difference as follows. In *Se*, there are five Ser tRNA genes of which one with CGA anticodon had 9 substitutions, one tRNA gene with GCT anticodon had 1 substitution, while two genes with GGA anticodon and one gene with TGA anticodon had 0 substitutions. In *Sa*, isoacceptor tRNA genes in case of Ser as well as Gly were observed to have different substitutions. There are five Ser tRNA genes of which one with GCT anticodon had no substitutions while one with GGA anticodon had 9 substitutions. There are seven Gly tRNA genes of which two tRNA genes with GCC anticodon had no substitutions but five Gly tRNA with TCC anticodon had 0 to 9 substitutions. In *Sp*, there are three Lys tRNA genes of which one with CTT anticodon and the other two with TTT anticodon. It is interesting that the tRNA gene with CTT anticodon was observed with 10 substitutions while the other tRNA genes with TTT anticodon were with no substitutions. Future studies will elucidate these differences observed among the isoacceptor tRNA genes.

5.4 Discussion

Secondary structure in transcripts is important for its function and intra-strand base pairing is important for their stability. Transfer RNA genes are known to have well defined secondary structures unlike IRs. Though it is known in literature that *ti* frequency is higher than *tv*, role of RNA secondary structure towards it has not been explored adequately at species level. Our endeavor in this aspect is to study *ti* and *tv* in tRNA genes and compare these substitutions between loop and stem regions. We have observed that in comparison with IRs, *tv* frequency is proportionately lower than that of *ti* in tRNA genes. This observation is in concordance with the assumption that secondary structure region is likely to have low *tv* frequency. Further we have compared *ti* and *tv* between loop and the stem regions. In stem regions *ti* were proportionately higher than *tv* when compared with the loop region. This is in concordance with the assumption made in this study that *tv* is more deleterious in the stem regions than *ti*. It is known that G:U is a more stable pair than A:C pair in tRNA stem. Therefore, *ti* substitution from amino base (A/C) to keto base (G/T) that results stable G:U pairing is likely to be preferred over the *ti* substitution from keto base (G/T) to amino base (A/C) that results unstable A:C pairing. In concordance to this hypothesis in tRNA stem region transition substitution from amino base (A/C) to keto base (G/T) is observed to be significantly higher than that in the loop region. This further supports the notion that secondary structure in tRNA influences base substitutions. It may be the postulation that *ti* and *tv* mutations are fixed/selected according to evolutionary constraints associated with maintenance of tRNA secondary structure.

Previous researchers had studied *ti* and *tv* in stem and loop regions of tRNA genes by comparing sequences across the species. They had observed more compensatory substitutions in tRNA stem region than non-compensatory substitutions. They had attributed the high *ti*/*tv* values in tRNA stem regions due to low frequency of compensatory *tv* in comparison to the frequency of compensatory *ti*. In the present work, we analyzed *ti* and *tv* in tRNA genes by comparing sequences within a species. So, we observed non-compensatory substitutions more in number than compensatory substitutions. It is likely that the non-compensatory substitutions are more recent ones in evolution than the compensatory substitutions. Non-compensatory mutations in tRNA genes stem regions are less likely to occur in nature because of its deleterious impact on the organism. Therefore, in the long run either the organism will have compensatory mutation or the mutation will be eliminated. This is why non-compensatory mutations are not found in interspecies tRNA gene comparison study [71]. However, the evolutionary distance of intraspecies variations are less and therefore these variations are expected to be more recent for which non-compensatory mutations are observed in this study. Among these substitutions, those having strong deleterious impact on tRNA structures are removed faster from the population in comparison to the ones that have moderate impact. It is assumed that *tv* are likely to have more deleterious impact on the secondary structure than *ti*. Therefore, we observed *tv* frequency is significantly lower than *ti* in stem region, in comparison to the *ti* and *tv* in loop regions. Further, *ti* substitutions with keto to amino results A:C pairing in the stem which destabilizes the secondary structure more than *ti* substitutions with amino to keto that results in G:U pairing. Though G:U pairing has been reported to be present in tRNA stem regions [75][78][130] there is recent report stating that, G:U pairing in tRNA genes having deleterious impact on mouse survivability [86]. Therefore, the former *ti* should be under stronger purifying selection than the latter. In concordance to it, amino to keto transition was observed to be higher than keto to amino transition in stem regions. However, no such pattern was observed in the loop region. Unlike the stem region, the loop region in tRNA undergoes several posttranscriptional modifications, that are important for their stability as well as various functions such as recognition by amino acyl tRNA synthetase, binding to translation factors, ribosome and codon anticodon recognition (decoding the codons) [3][4][147]. Therefore, any base substitutions arising at the modification site are likely to be selected against in the loop regions. This might be the reason for low substitution rates in the loop region. In future studies this sequence alignment studies might be helpful to find out residues in the loop region that undergoes post-transcriptional modifications. However, the influence of secondary structure on different base substitutions is evident from this analysis. Future research may be done to find out possible implications of compensatory as well as non-compensatory substitutions in tRNA gene in phylogenetic studies.

5.5 Conclusion

We believe our findings on mutations in stem vis-á-vis loop regions in tRNA genes might be extended to understand regions in mRNA that are potentially involved in secondary structure formation and gene expression regulation. It is known that the presence of secondary structure in mRNA is important in case of rho-independent

[2] and rho-dependent [98] transcription termination, pausing of ribosome [9] during translation and protein folding. It is also known that secondary structure near the ribosome binding site or Shine-Dalgarno sequence plays important role in translation regulation [66]. The role of RNA secondary structure on riboswitches is well documented in the literature [65][134]. Further, there are regions in other RNA such as miRNA [124] and introns in eukaryotes [109] involved in secondary structure formation. The difference between amino to keto and keto to amino transitions in the stem regions of tRNA might be of importance to understand the potential stem region in the transcript. In the stem regions of secondary structures, amino to keto is expected to be more compared to keto to amino transitions. Exploring this in future, might be helpful to understand the contribution of RNA secondary structure on gene regulation. Further, it has been reported that the GC content of the paired stem regions of the 16S rRNA genes positively correlates with the optimal growth temperature of bacteria and archaea [229]. Therefore, it might be anticipated that difference between the two transitions such as amino to keto and keto to amino will be more in thermophiles in comparison to mesophiles or psychrophiles. In future it will be interesting to study this in bacteria.