# Chapter 2

# Estimation Of ti/tv Ratio by Accounting Degeneracy and Pretermination Nature of Codons

# CHAPTER 2

# ESTIMATION OF TI/TV RATIO BY ACCOUNTING DEGENERACY AND PRETERMINATION NATURE OF CODONS

## 2.1. Abstract

Transition (*ti*) and transversion (*tv*) are the major causes for genome variation. The accurate estimation of *ti* to *tv* ratio ($\frac{ti}{tv}$) in genomes is crucial for understanding of mutational and selection processes in organisms as it is influenced by both codon degeneracy and pretermination codons (PTC). Therefore, we developed a method (accessible at https://github.com/CBBILAB/CBBI.git) to estimate $\frac{ti}{tv}$ ratio by accounting codon degeneracy as well as PTC in protein coding sequences. Our findings revealed a distinct impact of codon degeneracy and PTC on the $\frac{ti}{tv}$ ratio in the *Escherichia coli* genome. We observed a decreasing order among the frequencies of different base substitutions such as synonymous transition (*Sti*) > synonymous transversion (*Stv*) > non-synonymous transition (*Nti*) > non-synonymous transversion (*Ntv*) in *E. coli* genome. The correlation was strong between *Sti* and *Stv* values (Pearson *r* value 0.795) whereas the correlation was weak between *Sti* and *Nti* (Pearson *r* value 0.192). Coding sequences with similar *Sti* values exhibited a wide range of *Nti* values. This indicated the varying strength of purifying selection acting on the coding sequences. In concordance with the assumption, the genes having higher *Nti* values were observed with lower codon adaptation index (CAI) values than that of the genes having lower *Nti* values. Our approach is convenient to visualize the frequency of base substitution variation as well as selection in protein coding sequences. The proposed method is useful to

estimate different $\frac{ti}{tv}$ ratios accurately in coding sequences and is insightful from an evolutionary perspective.

## 2.2. Introduction

The degeneracy in the genetic code table is an important feature to study molecular evolution (Crick et al., 1961; Khorana et al., 1966). The different degenerate codons influence the transition (*ti*) to transversion (*tv*) ratio ($\frac{ti}{tv}$) as follows: synonymous variation in a *two-fold degenerate* (TFD) codon occurs *via* only one transition whereas the same in a *four-fold degenerate* (FFD) codon occurs via one transition and two transversions (Supplementary Fig. 1a). In addition to the above, compositional variation of pretermination codons (PTC) (Supplementary Fig. 1b and 1c) (Modiano et al., 1981) among the coding sequences is also likely to influence the $\frac{ti}{tv}$ ratio because non-synonymous variations leading to termination codons purged out rapidly in a population due to stronger purifying selection (Li et al., 1981; Eynden et al., 2016; Morales et al., 2021). Considering in general, a *ti* being more frequent than a *tv* (Seplyarskiy et al., 2012; Duchêne et al., 2015; Stoltzfus and Norris, 2016; Lewis et al., 2016; Lyons and Lauring, 2017; Schroeder et al., 2017; Sen et al. 2022) which is known as Kimura's two parameter model (Kimura 1980), and a stronger purifying selection on non-synonymous variation than a synonymous variation (Hurst and Pál 2001), the $\frac{ti}{tv}$ ratios of TFD codons and FFD codons are expected to be different. Accordingly, it has been recently reported that the synonymous variation in TFD codons is relatively higher than the FFD codons in *Escherichia coli* (Aziz et al., 2022). Therefore, in any organism, the $\frac{ti}{tv}$ ratio is likely to vary across the coding sequences having differences in degenerate codon composition (Beura et al., 2023). Hence, an estimation of $\frac{ti}{tv}$ ratio in a coding sequence by accounting codon degeneracy and PTC compositions will be important to further carrying out comparative analysis of this value among coding sequences in a genome.

Considering the number of possibilities of the two types of base variations in genomic DNA (Fersht and Knill-Jones 1981) (Supplementary Fig. 1d). The ratio of *ti* to *tv* is theoretically expected to be 0.5. But the ratio is generally observed more than 0.5 in the genomes (Duchêne et al., 2015; Stoltzfus and Norris, 2016). Because of the following factors such as geometry of DNA double helix (Topal and Fresco, 1976), modification of DNA bases such as deamination of cytosine and adenine (Kino and Sugiyama, 2001; Rocha et al., 2006; Bhagwat et al., 2016), secondary structure in the RNA (Sen et al., 2022) and codon degeneracy are known to elevate *ti* phenomenon in a genome over *tv* (Muse and Gaut, 1994; Aziz et al., 2022). In a standard genetic code table out of 549 total single nucleotide variations (SNVs) involving 61 sense codons (9 SNVs per codon), only 134 SNVs are synonymous (24.4%). Among synonymous (*Sti* and *Stv*) and non-synonymous (*Nti* and *Ntv*) variations, the theoretically estimated ($_e$) numbers of *Sti*, *Stv*, *Nti* and *Ntv* in a genetic code table are 62, 72, 121 and 294, respectively (Supplementary Table 1). Yet the observed synonymous variations are always higher than the expected proportion across genes (Wolfe et al., 1989; Tamura, 1992; Moriyama and Powell, 1997; Ngandu et al., 2008) due to stronger purifying selection on non-synonymous variations than the synonymous variations (Hurst and Pál, 2001). Hence, by accounting the codon degeneracy as well as PTC, the *Sti*, *Stv*, *Nti* and *Ntv* in the genetic code table can be estimated theoretically for each codon (Supplementary Table 2).

The codon substitution model suggests that sites within codons evolve at different rates and consequently they should not be equally treated (Muse and Gaut, 1994; Shapiro et al., 2006; Arenas, 2015). In the codon substitution model, using likelihood approach the $\frac{ti}{tv}$ ratio was estimated in accordance with the codon degeneracy class which was implemented in estimating *dN*/*dS* in coding sequences (Muse and Gaut, 1994; Goldman and Yang, 1994). Recently, by accounting $\frac{ti}{tv}$ substitution rate difference in codons as well as nonsense variations in PTC, the *dN*/*dS* values have been improved (Aziz et al., 2022). This finding implies that factors such as codon degeneracy and PTC composition should be considered while analyzing the evolution of protein-coding sequences. The ratio of *ti* to *tv* has been studied for reliable estimation of sequence distance and phylogeny reconstruction (Yang and Yoder, 1994). Using maximum likelihood approach, the $\frac{ti}{tv}$ ratio (commonly known as kappa $\kappa$) has been used to mainly compare the $\frac{ti}{tv}$ values among different species (Yang and Yoder, 1994). The parameter $\kappa$ has not been implemented to compare coding sequences within a genome in the context of their compositional differences of different degenerate classes before. As the selection pressure on synonymous and non-synonymous variations are heterogeneous (Bartolomé et al., 2005), it is important to analyze the $\frac{ti}{tv}$ values among different coding sequences within an organism by accounting synonymous sites and non-synonymous sites regarding their compositional differences of different degenerate codons.

It is noteworthy that in a genome, the number of potential replacements is greater for non-synonymous alterations, whereas the frequency of actual variations is higher for synonymous alterations among coding sequences. The proposed improved estimators account for the observed variations out of the total possible variations. On the contrary, the conventional method only takes account of the observed *ti* and *tv*. The possible sites for *Sti* and *Stv* are also different among coding sequences, which is dependent upon the codon composition belonging to different degeneracy classes. The conventional $\frac{ti}{tv}$ ratio does not adequately account for the limitations of variable selection against synonymous and non-synonymous changes. Given that there is a significant variation of substitution rates among protein coding sequences within a genome (Wu and Li, 1985), an appropriate methodology for the estimation of *ti* to *tv* ratio is required. We are presenting an improved estimator that builds upon the research conducted by Yasuo Ina in 1998 (Ina, 1998). Considering the differential selective constraint in synonymous and non-synonymous variations it is crucial to separately analyze the synonymous $\frac{ti}{tv}$ and non-synonymous $\frac{ti}{tv}$ to gain a comprehensive understanding of the overall $\frac{ti}{tv}$. Hence, all these parameters make the requirement of an improved estimator instead of the conventional estimator. The proposed estimators used in this study are helpful to estimate the *ti* and *tv* rate differences in coding sequences by accounting the compositional difference in terms of codon degeneracy and PTC. The proposed estimator is useful in comparative study of mutation bias between two coding sequences in an organism, as it normalizes the $\frac{ti}{tv}$ ratio in form of observed variations out of total possible variations. Further, we observed a correlation between TFD: FFD and $\frac{Sti}{Stv}$ values and the PTC% and $\frac{Nti}{Ntv}$ values. We tried to establish the biological significance of codon composition and $\frac{ti}{tv}$ ratio across 2516 coding sequences of *E. coli*. Interestingly, coding sequences with similar synonymous variation frequency are found to have a wide variation regarding their non-synonymous variation frequency. The improved method developed in the present study may be applied to study different $\frac{ti}{tv}$ ($\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$) across coding sequences.

## 2.3. Materials and Methods

### 2.3.1. Coding sequences of *E. coli* genome and finding out *ti* and *tv*

In this study, we carried out a computational analysis of SNVs in 2516 protein coding sequences across 157 strains of *E. coli* (Thrope et al., 2017). The information regarding the dataset selected for the present study is provided in Supplementary Table 3. Although the public database has

~4500 coding sequences in *E. coli* genome, we excluded those coding sequences exhibiting size differences and improper alignment/annotation across the strains. Codons having double or triple substitutions or variations represented by an ambiguous nucleotide were not considered further in the analyses of the variations. The procedure in detail used for estimating SNVs is represented in Supplementary Table 4 using example of a hypothetical sequence. This procedure is based on intra-species genome sequence comparison to find SNVs in bacterial genomes. A reference sequence was generated considering the most frequent nucleotide occurrence at a site in the gene sequence. The logic behind the derivation of the reference sequence is that the most abundant nucleotide in a position is considered as the ancestral nucleotide for the position (Sen et al., 2022; Aziz et al., 2022; Sen, 2022, Beura et al., 2023).The observed SNVs ($SNV_o$) at different positions of each codon were categorized into $Sti_o$, $Stv_o$, $Nti_o$ and $Ntv_o$, where S for synonymous, N for non-synonymous, $ti_o$ for observed transition and $tv_o$ for observed transversion, and subsequently $ti_o$ ($Sti_o$ $_+$ $Nti_o$) and $tv_o$ ($Stv_{o+}$ $Ntv_o$) for each gene sequence were calculated ( Supplementary Table 5). After finding out the observed variation values for a gene, we estimated the total number of variations such as $Sti_e$, $Stv_e$ $Nti_e$, $Ntv_e$, $ti_e$ and $tv_e$ possible theoretically in the gene. We derived the reference sequence for each gene based on the methodology explained above. The reference sequence was used for the theoretical estimation of these variations. For example, if GGG codon abundance is 10 in a gene, then the total theoretically estimated value for *Sti*, *Stv*, *Nti* and *Ntv* are 10, 20, 20 and 40 (*Sti*, *Stv*, *Nti* and *Ntv* for GGG codon is 1, 2, 2, and 4, respectively). The *Sti*, *Stv*, *Nti*, and *Ntv* for each codon in the genetic code table are given in Supplementary Figs (1a, 1b, 1c). This information for each codon was used to calculate the theoretically estimated SNVs ($Sti_e$, $Stv_e$, $Nti_e$ and $Ntv_e$) for each gene under the study (Supplementary Table 6). The values of the observed SNVs ($Sti_o$, $Stv_o$, $Nti_o$, $Ntv_o$, $ti_o$, and $tv_o$) and the values of estimated SNVs ($Sti_e$, $Stv_e$ $Nti_e$, $Ntv_e$, $ti_e$, and

$tv_e$) of a coding sequence are then used in the modified estimator to find out different $\frac{ti}{tv}$ ratio ($\frac{Sti}{Stv}$; $\frac{Nti}{Ntv}$) (Supplementary Table 7). A detailed workflow is provided on Fig. 2.1.

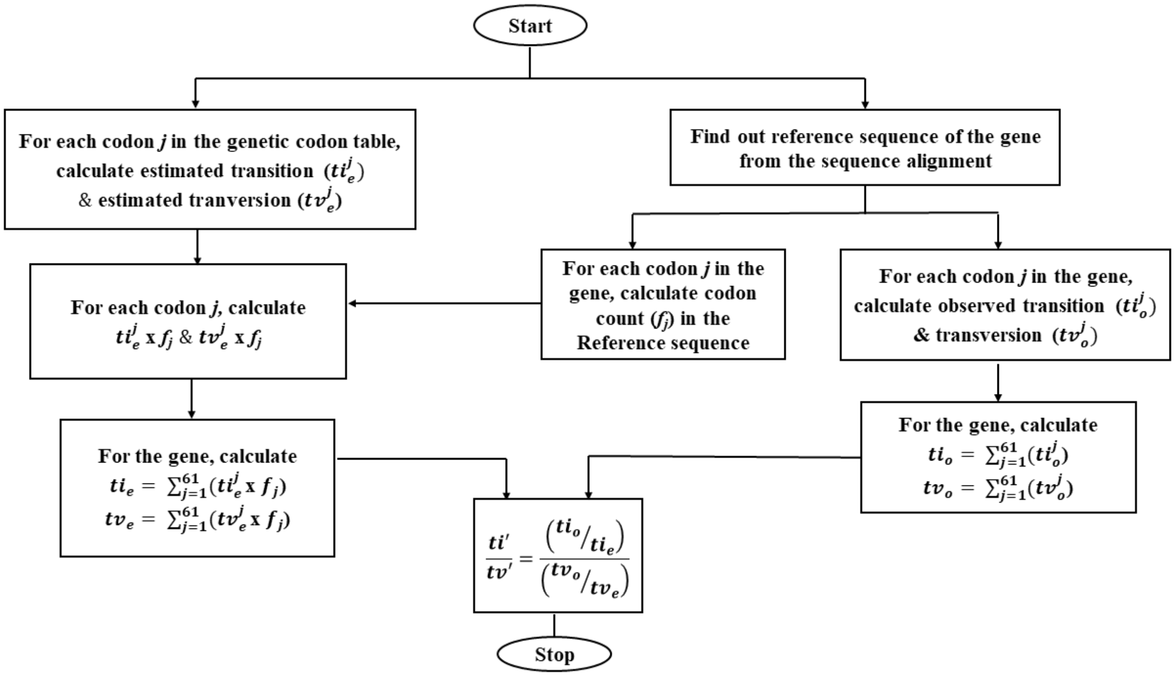

Start
For each codon $j$ in the genetic codon table, calculate estimated transition $(ti_e^j)$ & estimated tranversion $(tv_e^j)$
Find out reference sequence of the gene from the sequence alignment
For each codon $j$, calculate $ti_e^j$ x $f_j$ & $tv_e^j$ x $f_j$
For each codon $j$ in the gene, calculate codon count ($f_j$) in the Reference sequence
For each codon $j$ in the gene, calculate observed transition $(ti_o^j)$ & transversion $(tv_o^j)$
For the gene, calculate $ti_e = \sum_{j=1}^{61}(ti_e^j \text{ x } f_j)$ $tv_e = \sum_{j=1}^{61}(tv_e^j \text{ x } f_j)$
For the gene, calculate $ti_o = \sum_{j=1}^{61}(ti_o^j)$ $tv_o = \sum_{j=1}^{61}(tv_o^j)$
$\frac{ti'}{tv'} = \frac{\left({ti_o}/{ti_e}\right)}{\left({tv_o}/{tv_e}\right)}$
Stop

**Fig. 2.1.** Schematic representation demonstrates step wise workflow for calculation of $\frac{ti'}{tv'}$ using the improved method. Step wise calculations for the workflow are given for a sample gene in the Supplementary Table S5, S6, S7.

### 2.3.2. Improved estimators

We proposed an improved mathematical equation for the calculation of different $\frac{ti}{tv}$ ratio as described below.

$$\frac{ti'}{tv'} = \frac{\left({}^{ti_o}/_{ti_e}\right)}{\left({}^{tv_o}/_{tv_e}\right)} \qquad \textbf{Equation 1}$$

$$\frac{Sti'}{Stv'} = \frac{\left({}^{Sti_o}/_{Sti_e}\right)}{\left({}^{Stv_o}/_{Stv_e}\right)} \qquad \textbf{Equation 2}$$

$$\frac{Nti'}{Ntv'} = \frac{\left({}^{Nti_o}/_{Nti_e}\right)}{\left({}^{Ntv_o}/_{Ntv_e}\right)} \qquad \textbf{Equation 3}$$

Where, $\frac{ti'}{tv'}$ is improved $\frac{ti}{tv}$ ratio, $\frac{Sti'}{Stv'}$ is improved $\frac{Sti}{Stv}$ ratio, $\frac{Nti'}{Ntv'}$ is improved $\frac{Nti}{Ntv}$ ratio: $ti_o$ is number of transitions observed, $\boldsymbol{ti_e}$ is the number of possible transitions estimated theoretically, $tv_o$ is the number of transversions observed, $tv_e$ is the number of possible transversions estimated theoretically, $\boldsymbol{Sti_o}$ is number of synonymous transitions observed, $\boldsymbol{Sti_e}$ is the number of theoretically estimated synonymous transitions, $\boldsymbol{Stv_o}$ is number of synonymous transversions observed, $\boldsymbol{Stv_e}$ is the number of theoretically estimated synonymous transversions, $\boldsymbol{Nti_o}$ is number of non-synonymous transitions observed, $\boldsymbol{Nti_e}$ is the number of theoretically estimated non-synonymous transitions, $\boldsymbol{Ntv_o}$ is number of non-synonymous transversions observed, $\boldsymbol{Ntv_e}$ is the number of theoretically estimated non-synonymous transversions. We calculated $\frac{ti}{tv}$ ratios for all the coding sequences using a program: a script written in Python-language is available at GitHub in the following link: (https://github.com/CBBILAB/CBBI.git).

Using our improved approach, the estimated *Sti, Stv, Nti*, and *Ntv* values for the 2516 coding sequences were calculated and analyzed. To further understand the better applicability of our approach, we found out $\frac{ti}{tv}$ in all these 2516 coding sequences using maximum likelihood (ML) approach available in MEGA-X (Kumar et al., 2018) and compared it with the values estimated using the improved approach described in this study.

### 2.3.3. Comparative representation of different $\frac{ti}{tv}$ through the conventional and the improved estimators

A comparative analysis of $\frac{ti}{tv}$ values calculated by using the conventional method and the improved method was performed to prove the better accuracy of our method proposed in this study. Also, we calculated the ratio between the *two-fold degenerate* and the *four-fold degenerate* codons (TFD: FFD) and the percentage of pretermination codons (PTC%) in each gene and performed statistical analysis between different parameters. The percentage (%) change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values obtained by conventional and improved methods calculated as follows.

(a) $$\frac{Sti}{Stv}\ \%\ \text{change} = \frac{\frac{Sti}{Stv} - \frac{Sti'}{Stv'}}{\frac{Sti}{Stv}} \times 100$$

(b) $$\frac{Nti}{Ntv}\ \%\ \text{change} = \frac{\frac{Nti}{Ntv} - \frac{Nti'}{Ntv'}}{\frac{Nti}{Ntv}} \times 100$$

### 2.3.4. Statistical analyses

OriginPro, Version 2022, OriginLab Corporation, Northampton, MA, USA was used to draw the Box-plot/Scatter plots as well as to perform the Mann-Whitney test (Mann and Whitney, 1947) to find out the *p*-value. The correlation plot was drawn for $R^2$ value and Pearson's correlation coefficient (Pearson, 1896) was also calculated between the $\frac{ti}{tv}$ values obtained using the conventional and improved method.

## 2.4. Results

### 2.4.1. Impact of codon degeneracy and PTC on $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values in *E. coli*

We found out the total number of TFD, FFD, and PTC codons across 2516 coding sequences of *E. coli* and the percentage of TFD, FFD, and PTC was calculated (Supplementary Table 3). The percentage (%) compositional differences in the 2516 coding sequences represented as TFD%, FFD%, and PTC% were presented using box-plot (Supplementary Fig. 5). A variable range of TFD: FFD ratio was observed in the coding sequences of *E. coli.* The minimum to maximum TFD:FFD ratio was observed as 0.229 and 3.333 in *sugE* and *mntS*, respectively. A summary of the different parameters observed throughout the dataset was presented in Table 2.1 Among 2516 coding sequences, only 35 coding sequences possessed the TFD:FFD ratio of 1.000, which suggested that there exists a compositional asymmetry between TFD codon and FFD codon compositions across the coding sequences. We presented the TFD%, FFD% and PTC%, using a multi-paneled scatter plot to study the correlation between any two pairs (Fig. 2.2). As expected, the FFD and the TFD compositions exhibited a negative correlation (Pearson *r* value -0.56). The FFD and the PTC compositions exhibited a negative correlation (Pearson *r* value -0.65), while the TFD and the PTC compositions exhibited a positive correlation (Pearson *r* value 0.83) because the eighteen PTC codons include ten TFD codons and only one FFD codon. It also explained a positive correlation (Pearson *r* value 0.81) observed between TFD: FFD ratio and the PTC. The compositional variation (TFD, FFD and PTC) observed in coding sequences were considered to validate the accuracy of estimation of different $\frac{\boldsymbol{ti}}{\boldsymbol{tv}}\ (\frac{Sti}{Stv}; \frac{Nti}{Ntv})$ ratio using conventional and the improved method.

**Table 2.1.** Nucleotide and codon compositional features of the 2516 coding sequences considered in the study

| Genes considered | Size bp (Min-Max) | GC% (Min-Max) | TFD/FFD (Min-Max) | PTC% (Min-Max) | TFD% | FFD% | PTC% | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 33.384 | 34.084 | 36.350 | Mean |
| 2516 | 93-4554 | 32.95-63.80 | 0.22-3.33 | 15.00-57.14 | 5.726 | 4.761 | 5.625 | S. D |

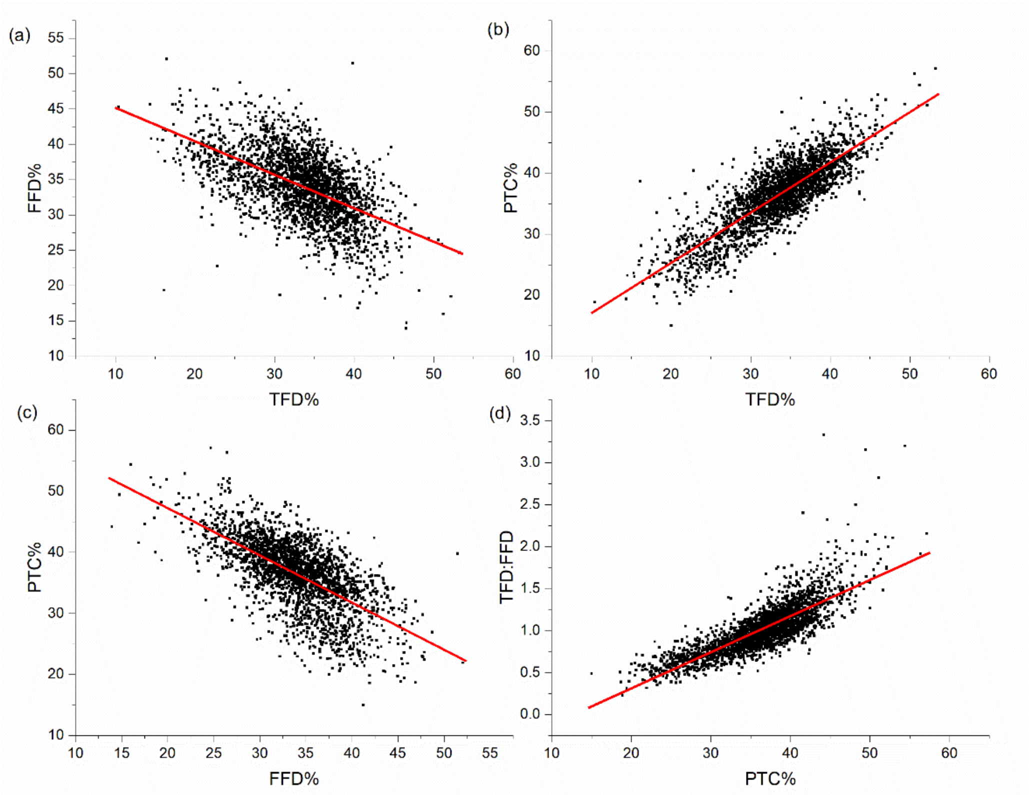


**Fig. 2.2.** The multi-paneled scatter plots elucidate the comparison of codon composition between different parameters such as TFD% to FFD%, FFD% to PTC%, TFD% to PTC % and TFD: FFD to PTC%. In total 2516 *E. coli* genes were considered while calculating these values. The *x*-axes and the *y*-axes represent different parameters of codon compositions in the individual plots. The Pearson *r* (TFD%, FFD%) and Pearson *r*(FFD%, PTC%) were observed to be -0.56 and -0.65 respectively, indicating a negative moderate correlation. Whereas the Pearson *r* (TFD%, PTC%) and Pearson *r* (TFD: FFD, PTC%) were observed to be 0.83 and 0.81 respectively, indicating a strong correlation.

We estimated $\frac{ti}{tv}$ by accounting codon degeneracy as well as PTC in coding sequences and compared with the conventional method of $\frac{ti}{tv}$ in coding sequences. Theoretically, a coding region entirely composed of FFD codons will result into a twice $\frac{Sti'}{Stv'}$ value than $\frac{Sti}{Stv}$ value, similarly a coding region composed of same proportion of FFD and TFD codons will result into equal values of $\frac{Sti'}{Stv'}$ and $\frac{Sti}{Stv}$. The increase in TFD codon proportion will have direct impact on the difference between $\frac{Sti}{Stv}$ values and $\frac{Sti'}{Stv'}$ values. This pattern was observed in a hypothetical sequence considered as presented in the Supplementary Fig. 2.

The complete set of observed and estimated values of different variations in the 2516 coding sequences of *E. coli* were enlisted (Supplementary Table 3). These values were used to estimate the $\frac{ti}{tv}, \frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ using the conventional method. The improved method was used to also estimated the $\frac{ti'}{tv'}, \frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$. A comparative analysis was made between the values derived using the conventional method ( $\frac{ti}{tv}, \frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ ) and the values derived using improved method ($\frac{ti'}{tv'}, \frac{Sti'}{Stv'}$ and

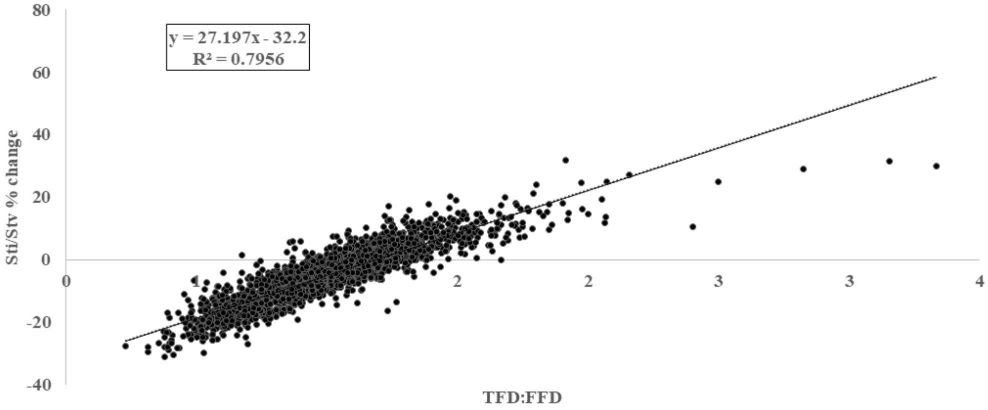
$\frac{Nti'}{Ntv'}$ ) (Supplementary Table 3). A box-plot showing the comparison between the values of different $\frac{ti}{tv}$ ratio obtained through the conventional and improved method is represented in Supplementary Fig. 7a. It was evident that the values estimated using the improved method were significantly different from that estimated using the conventional method ($p$ value <0.01). The change in conventional and improved ratios between the synonymous and the non-synonymous $\frac{ti}{tv}$ could be better visualized in Supplementary fig. 7b.

The synonymous variation in the coding sequences is likely to be different considering the differential composition of TFD and FFD codons, because when FFD codon proportion is higher in a gene, $\frac{ti}{tv}$ value is going to be lower. Similarly, when the FFD codon proportion is lower, $\frac{ti}{tv}$ value is going to be higher. We analyzed the coding sequences exhibiting significant discrepancies between the two set of values ($\frac{Sti}{Stv}$, $\frac{Sti'}{Stv'}$). Some of the genes exhibiting higher value in the $\frac{Sti'}{Stv'}$ are DNA binding transcriptional regulation, ABC transporter, DNA polymerase III subunits, and Cytochrome Bo3 subunits. These coding sequences are comprised of higher FFD% than the TFD%. Some of the genes such as transcriptional regulator, toxin-antitoxin biofilm protein, and Iron-Sulphur cluster exhibited lower value in the $\frac{Sti'}{Stv'}$ as these coding sequences are composed of higher TFD% than FFD %.

The Pearson correlation coefficient value between TFD: FFD and % change in $\frac{Sti}{Stv}$ (conventional to improved) was obtained to be 0.891, signifying a strong correlation. Similarly, the Pearson correlation coefficient value between PTC% and % change in $\frac{Nti}{Ntv}$ was obtained to be -0.488, a moderate negative correlation between the two variables. The results were found to be statistically significant at $p$<0.01, for $\frac{ti}{tv}$ & $\frac{ti'}{tv'}$, $\frac{Sti}{Stv}$ & $\frac{Sti'}{Stv'}$ and $\frac{Nti}{Ntv}$ & $\frac{Nti'}{Ntv'}$. Further the correlation graph between %

change in $\frac{\boldsymbol{Sti}}{\boldsymbol{Stv}}$ and TFD: FFD and % change in $\frac{\boldsymbol{Nti}}{\boldsymbol{Ntv}}$ and PTC% also substantiate the role of codon composition and $\frac{\boldsymbol{ti}}{\boldsymbol{tv}}$ ratio (Fig. 2.3a & 2.3b). Further, we performed a simulation study to substantiate the impact of the compositional difference of degeneracy class in estimating $\frac{\boldsymbol{Sti}}{\boldsymbol{Stv}}$ by using our improved method in coding sequences (Supplementary Table 8, Supplementary Fig. 3). The values obtained by the improved method were also different from the values obtained by the ML based method (Supplementary Table 9, Supplementary Fig. 4), as the degeneracy as well as PTC was not accounted in the ML based method.

(a)



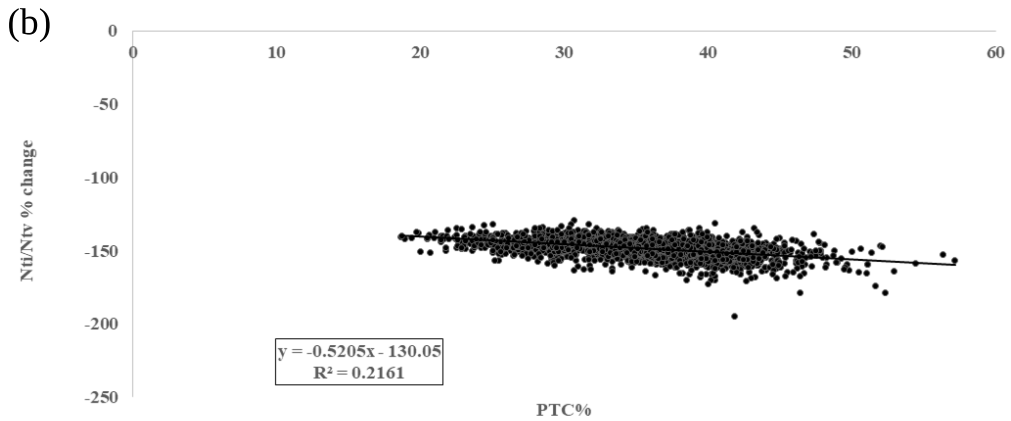


**Fig. 2.3.** Regression plots between % change in $\frac{Sti}{Stv}$ and TFD: FFD is presented in fig. 2.3.(a) and that between $\frac{Nti}{Ntv}$ and PTC% is presented in fig. 2.3(b) for the *E. coli* genes. Percentage change in $\frac{Sti}{Stv}$ is showing a strong correlation with the TFD: FFD with a Pearson correlation coefficient value 0.89, whereas % change in $\frac{Nti}{Ntv}$ is showing a weak negative correlation with the PTC% with a Pearson correlation coefficient value -0.46.

### 2.4.2. *Sti*, *Stv*, *Nti* and *Ntv* frequency in coding sequences

The frequency of *Sti*, *Stv*, *Nti* and *Ntv* were calculated and compared across the 2516 coding sequences. For Instance, the *Sti* frequency was calculated by considering the number of total *Sti* changes divisible by total possible *Sti* changes in a gene. Similarly, the remaining parameters (*Stv*, *Nti* and *Ntv*) were also calculated using the above logic, which follows our improved estimator i.e. observed number of changes out of total possible changes. The Supplementary Table 11 represents the overall *Sti*, *Stv*, *Nti* and *Ntv* values of the 2516 coding sequences considered for the study. The mean frequency values of *Sti*, *Stv*, *Nti*, and *Ntv* in the 2516 coding sequences were 0.113, 0.034, 0.014 and 0.004, respectively. (Supplementary Table 11, Fig. 2.4). This suggested the decreasing order of the different substitutions in genome such as *Sti* > *Stv* >*Nti* > *Ntv*. The values suggested that *Sti* was

3.3 times more frequent than *Stv*, *Nti* was 3.7 times more frequent than *Ntv*, *Sti* is 8.2 times more frequent than *Nti* and *Stv* was 9.3 times more frequent than *Ntv*. The relative purifying selection on *Nti* and *Ntv* in relation to their synonymous counter parts were similar. *Sti*, the most frequent substitution, was 28.25 times more frequent than *Ntv*, the least frequent substitution in this study. This we believe a distinct demonstration of frequencies of different substitutions such as *Sti*, *Stv*, *Nti*, and *Ntv* in *E. coli* coding regions.

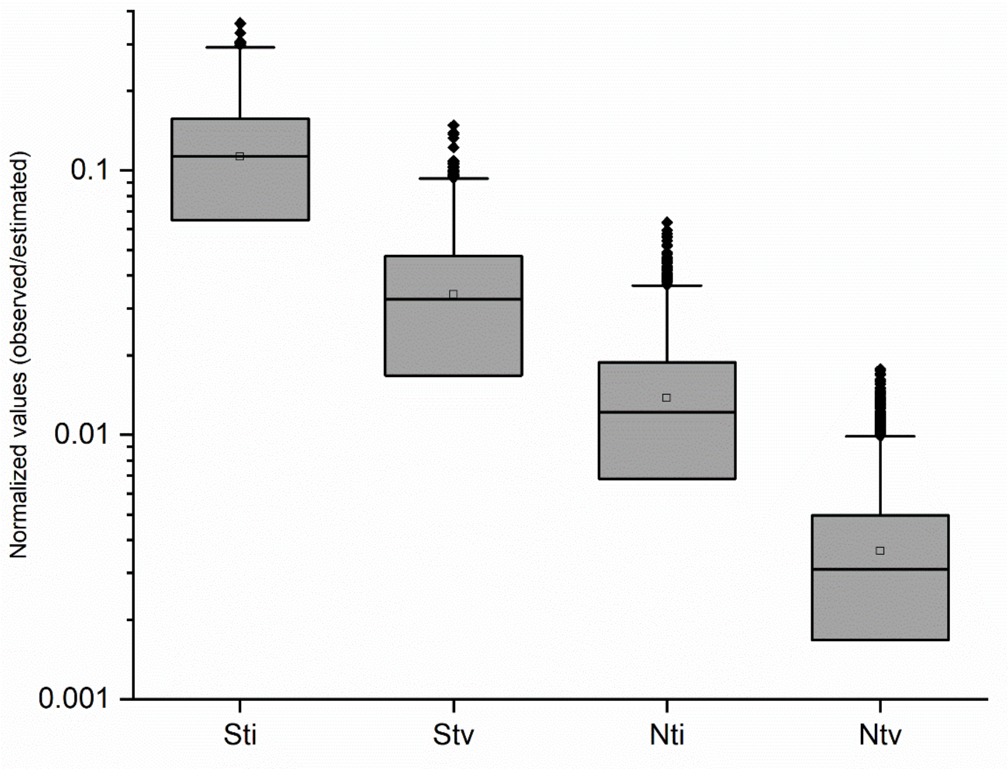


**Fig. 2.4.** The figure illustrates distribution of *Sti*, *Stv*, *Nti*, and *Ntv* values using box-plot. The log scale has been used in the *y*-axis. The mean *Sti* and *Stv* were observed as 0.113 and 0.031 respectively, similarly the mean *Nti* and *Ntv* values were observed as 0.014 and 0.004. *Sti* was observed to be more than *Stv* and *Nti* was observed to be more than *Ntv*. The *Sti* and *Stv* were significantly different ($p<0.01$) from *Nti* and *Ntv* respectively.

Further we did a Pearson correlation study between different base substitution values that were obtained by using the conventional estimator as well as by the improved estimator (Table 2.2). The Pearson correlation value (*r*) between *Sti* and *Stv* was 0.920 in conventional method and 0.795 in improved method, respectively. Between *Nti* and *Ntv*, the Pearson correlation value was observed as 0.789 in conventional method and 0.612 in improved method, respectively. A lower correlation in the case of improved method was expected considering the normalization of coding region size in this method. Therefore, the correlation result using the values from the improved method was the true representation than the correlation values obtained in case of the conventional method where size of the coding region was not normalized. This was more evident in the correlation between *Sti* and *Nti*: the Pearson correlation value (*r*) between *Sti* and *Nti* was 0.527 in conventional method whereas the same was 0.192 in case of the improved method. The low correlation is expected here due to the independent nature of the purifying selection on non-synonymous variation. Therefore, our improved method is more appropriate in estimating the distinctive selection patterns on coding regions by doing comparison between *Sti* and *Nti* (Table 2.2).

**Table 2.2.** Pairwise Pearson correlation coefficient among *Sti*, *Stv*, *Nti* and *Ntv* through conventional and improved estimator

| Pearson correlation coefficient | Conventional method | Improved method |
|---|---|---|
| | Observed values | Normalized values (Obs/Est) |
| *Sti & Stv* | 0.920 | 0.795 |
| *Nti & Ntv* | 0.789 | 0.612 |
| *Sti & Nti* | 0.527 | 0.192 |
| *Stv & Ntv* | 0.544 | 0.257 |

As part of our comparative study, we investigated the coding sequences of similar *Sti* and different *Nti* values (Fig. 2.5). In a set of 20 coding sequences with similar *Sti* values such as 0.131, the *Nti* values were ranging from 0.000 to 0.030. This wide range of *Nti* values, with a maximum 30-

fold variation, highlighted the presence of a variable *Nti* pattern within the respective coding sequences. Similar observation could be found in other *Sti* values. We compared *Sti* and *Nti* values in relation to gene expression represented with codon adaptation index (CAI) values (Sharp and Li 1987; Sen et al. 2019). We studied coding sequences with size more than 500 bp. The top 100 coding sequences with maximum *Sti* values (mean 0.255) were compared with bottom 100 coding sequences with minimum *Sti* values (mean 0.006). The mean CAI values for the top 100 coding sequences were 0.565 whereas the mean CAI values for the bottom 100 coding sequences was 0.509. The two mean CAI values were found to be close. This indicates the *Sti* values might not be influenced significantly due to gene expression. Similarly, the top 100 coding sequences with maximum *Nti* values (mean 0.036) were compared with bottom 100 coding sequences with minimum *Nti* values (mean 0.001). The mean CAI values for the top 100 coding sequences were 0.491 whereas the mean CAI values for the bottom 100 coding sequences was 0.638. The two mean CAI values of the coding sequences were found to be significantly different ($p$<0.01). The selection due to gene expression is stronger on *Nti* than on *Sti* in *E. coli*.

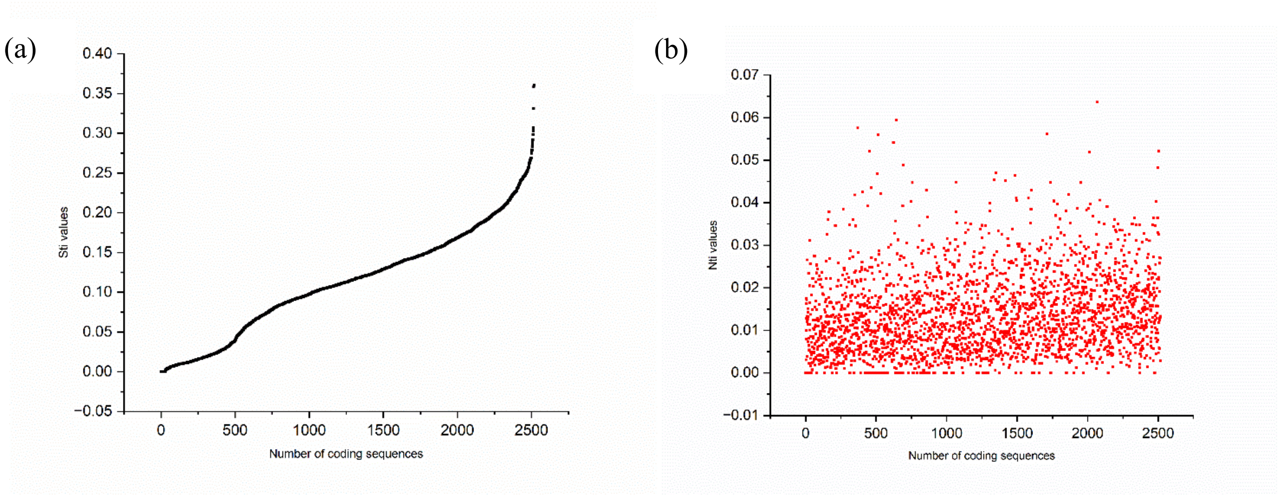


**Fig. 2.5.** The figure illustrates the *Sti* and *Nti* values of coding sequences used in the study through two separate plots. (a) The *y*-axis shows the *Sti* values of coding sequences while the *x*-axis shows the

number of coding sequences considered in this study. (b) The *y*-axis shows the *Nti* values of coding sequences while the *x*-axis shows the number of coding sequences considered in this study. We have sorted the *Sti* values in an ascending order, and the corresponding *Nti* values have been arranged in the same order accordingly. The black line (dots) represents the individual *Sti* values, and the red dots represent the corresponding *Nti* values in a coding sequence in the corresponding figures. The unidentical patterns in both the figures is an evident that many coding sequences with similar *Sti* values have a wide range of *Nti* values.

## 2.5. Discussion

Codon degeneracy influences the *ti* to *tv* ratio in coding sequences because a TFD codon undergoes synonymous variation only by transition while a FFD codon undergoes synonymous variation by transition as well as transversion (Aziz *et al.* 2022). In addition, there are only 18 pretermination codons in the genetic code table where nonsense substitution is rare to observe. Coding sequences in a genome vary with composition of codons belonging to different degeneracies as well as PTC. Therefore, it is important to compare coding sequences regarding their $\frac{ti}{tv}$ ratio by accounting degeneracy and PTC. Accordingly, in this study we have developed a methodology to estimate $\frac{ti}{tv}$ ratio by accounting to the above features in coding sequences. In total 2516 coding sequences of *E. coli* have been analyzed. The impact of codon degeneracy and PTC is observed distinctly on the ti/tv values in coding sequences. Interestingly, the coding sequences are observed to be different regarding their $\frac{ti}{tv}$ ratio even after accounting to the composition of degenerate codons and PTC. In fact, the $\frac{Sti}{Stv}$ values are observed to be variable among the coding sequences, which is intriguing. We did not observe any significant correlation of $\frac{Sti}{Stv}$ values with the gene expression represented as CAI values. The role of context dependent mutation (Sung et al., 2015; Zhu et al., 2017; Aikens et al., 2019; Ling et al., 2020), which has been described recently, might be a contributing factor that needs to be investigated in the future.

Our investigation into *ti* and *tv* revealed that *Sti* is around three times more frequent than *Stv* while it is eight times more frequent than *Nti*. The difference between *Sti* and *Stv* may be attributed due to the structural changes in DNA happening due to mispairing between either purine: purine or pyrimidine: pyrimidine but the difference between *Sti* and *Nti* may be attributed to the purifying selection acting on the coding sequences because of the non-synonymous changes. The magnitude of the purifying selection acting on non-synonymous substitution in *E. coli* genome is vividly observed. The strength of purifying selection on *Nti* and *Ntv* seems comparable considering the closeness of the $\frac{\boldsymbol{Sti}}{\boldsymbol{Stv}}$ and the $\frac{\boldsymbol{Nti}}{\boldsymbol{Ntv}}$ values. In a recent study by Zou and Zhang (Zou and Zhang, 2021) suggests that the purifying selection is more on *Ntv* than on *Nti* which is unlike the observation reported in our study. Our different observation from that of Zou and Zhang might be attributed to the intra-species study limited to *E. coli* whereas Zou and Zhang worked on an inter-species approach covering 90 clades representing all domains of life. Further, in our estimation of *Nti* and *Ntv* we have accounted codon degeneracy and PTC.

The correlation values between different substitutions such as *Sti* and *Stv*, as well as between *Sti* and *Nti* were higher in case of the conventional method than that in the case of the improved method. This is an important observation in our study and distinctly demonstrate the superiority of the improved method over the conventional method, which is explained as follows. In case there is no selection on *Sti*, the number of *Sti* is likely to be directly proportional with the number of synonymous transition sites available in a coding sequence. So, *Sti* is likely to correlate strongly with gene size. In case of *Nti*, purifying selection is strong, for which the number of *Nti* observed in a coding sequence is significantly lower than the number of *Sti* observed in the sequence though the number of theoretically possible *Nti* sites is always higher than the number of theoretically possible *Sti* sites. The number of observed *Nti* is also likely to be proportional with the possible *Nti* sites in a coding sequence. So, the number of *Nti* is likely to be proportional with gene size. Therefore, we can anticipate a strong positive corelation between *Sti* and *Nti* in the 2516 coding sequences. This is indeed what we have observed in our analysis here. In the improved method to estimate *Sti* and *Nti* described in this study has normalized the number of possible sites for *Sti* as well as *Nti* (or gene size). The improved *Sti* approach provides us an idea regarding the frequency of the substitution per *Sti* site whereas the improved *Nti* approach provides us an idea regarding the frequency of the substitution per *Nti* site. As size has been normalized, corelation between improved *Sti* and improved *Nti* is likely to be lower. Further, the value of improved *Sti*/improved *Nti* is more suitable to understand the purifying selection on a gene than *Sti*/*Nti*. A similar explanation may be also given for the *Sti* and *Stv* correlation. As gene size increases, *Sti* and *Stv* also increases, for which strong positive observed between the between the two values. In case of improved *Sti* and improved *Stv* cases the possible sites have been normalized respectively. Susceptibility of different sites for *ti* and *tv* are not same as evident by C→T *ti* and G→T *tv* are more frequent than the other *ti* and *tv*, respectively (Sen et al., 2022; Beura et al., 2023). Therefore, the correlation value between improved *Sti* and improved *Stv* is likely to be weaker than the same between *Sti* and *Stv* values.

As transitions are more frequent events than transversion, transition can be considered alone to understand the selection on coding sequences. Considering *Sti* representing the variation frequency in a coding sequence, the *Nti* representing the strength of purifying selection acting on the coding sequence, a comparison between *Sti* and *Nti* in a coding sequence might represent the strength of purifying selection acting on it. In several instances, we have observed that coding sequences with equal *Sti* variation frequency are highly different regarding their *Nti* variation frequency, which indicates disproportionate purifying selection on coding sequences. Additionally, when we compared the mean CAI values of the coding sequences with higher *Sti* values against the mean CAI values of the coding sequences with lower *Sti* values, we did not find any significant difference. On the contrary, when we conducted the same analysis for the *Nti* values, we observed a significant distinction between the CAI values of coding sequences with higher *Nti* values and those with lower *Nti* values. This observation indicates that highly expressed genes tend to have lower *Nti* values. Consequently, it suggests that purifying selection due to gene expression is more pronounced for *Nti* values than for *Sti* values. Since our study on selection relies on intra-species analysis, we anticipate that future research will unveil the underlying mechanisms of selection governing variation frequency in inter-species studies as well. A probable discrepancy between the process of selection between intra-species and inter-species can be drawn. This is a simple demonstration of observing purifying selection acting on coding sequences. Though the difference between *ti* and *tv* have been known for a long time, we could use it to study selection on coding sequences because of the improved estimator.

Various methods have been implemented in the estimation of the *ti* to *tv* ratio by molecular evolutionary biologists to understand the mutation as well as selection bias. The phylogeny-based method and Bayesian methods have been implemented by researchers in the inter-species study of $\frac{\boldsymbol{ti}}{\boldsymbol{tv}}$ (Purvis and Bromham, 1997; Huelsenbeck et al., 2001). Usually the *dN*/*dS* approach has been used to visualize the selection in coding sequences (Aziz et al., 2022) that considers both *ti* and *tv*. However, our method implemented in this study is simple, succinct, easy to follow for researchers in the field with limitations in mathematics and provides an insight to understand the evolution of protein coding sequences. We also did a comparison with the ML method. It has been said that the likelihood methods are computer intensive and difficult to apply to large datasets and can fall into local traps (Golding and Felsenstein, 1990). Furthermore, the extrapolation of our method in various inter-species studies, as well as phylogenetic tools can enhance the horizon of our understanding regarding mutational bias in coding sequences.

## 2. 6. Data and software availability

A script written in Python-language is available at GitHub in the following link: (https://github.com/CBBILAB/CBBI.git). Further queries regarding the software may contact SSS (ssankar@tezu.ernet.in). The authors confirm that the data supporting the findings of this study are available within the article and Supporting information.

## 2.7. Bibliography

- Aikens, R. C., Johnson, K. E., & Voight, B. F. (2019). Signals of variation in human mutation rate at multiple levels of sequence context. *Molecular Biology and Evolution*, *36*(5), 955-965.
- Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in genetics*, *6*, 163122.
- Aziz, R., Sen, P., Beura, P. K., Das, S., Tula, D., Dash, M., ... & Ray, S. K. (2022). Incorporation of transition to transversion ratio and nonsense mutations, improves the estimation of the number of synonymous and non-synonymous sites in codons. *DNA Research*, *29*(4), dsac023.
- Bartolomé, C., Maside, X., Yi, S., Grant, A. L., & Charlesworth, B. (2005). Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. *Genetics*, *169*(3), 1495-1507.
- Beura, P. K., Sen, P., Aziz, R., Satapathy, S. S., & Ray, S. K. (2023). Transcribed intergenic regions exhibit a lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of Escherichia coli and Salmonella enterica. *Journal of Genetics*, *102*(1), 22.
- Beura, P. K., Sen, P., Aziz, R., Chetia, C., Dash, M., Satapathy, S. S., & Ray, S. K. (2023). Difference in synonymous polymorphism related to codon degeneracy between cotranscribed genes in the genome of Escherichia coli. *Curr. Sci*.
- Bhagwat, A. S., Hao, W., Townes, J. P., Lee, H., Tang, H., & Foster, P. L. (2016). Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. *Proceedings of the National Academy of Sciences*, *113*(8), 2176-2181.

- Crick, F., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins.
- Duchêne, S., Ho, S. Y., & Holmes, E. C. (2015). Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC evolutionary biology*, *15*, 1-10.
- Fersht, A. R., & Knill-Jones, J. W. (1981). DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine. purine, purine. pyrimidine, and pyrimidine. pyrimidine mismatches during DNA replication. *Proceedings of the National Academy of Sciences*, *78*(7), 4251-4255.
- Golding, B., & Felsenstein, J. (1990). A maximum likelihood approach to the detection of selection from a phylogeny. *Journal of molecular evolution*, *31*, 511-523.
- Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, *11*(5), 725-736.
- Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755
- Hurst, L. D., & Pál, C. (2001). Evidence for purifying selection acting on silent sites in BRCA1. *TRENDS in Genetics*, *17*(2), 62-65.
- Ina, Y. (1998). Estimation of the transition/transversion ratio. *Journal of molecular evolution*, *46*(5), 521-533.
- Khorana, H. G., Büuchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kössel, H., ... & Wells, R. D. (1966, January). Polynucleotide synthesis and the genetic code. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 31, pp. 39-49). Cold Spring Harbor Laboratory Press.

- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*, 111-120.
- Kino, K., & Sugiyama, H. (2001). Possible cause of G· C→ C· G transversion mutation by guanine oxidation product, imidazolone. *Chemistry & biology*, *8*(4), 369-378.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, *35*(6), 1547.
- Lewis Jr, C. A., Crayle, J., Zhou, S., Swanstrom, R., & Wolfenden, R. (2016). Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences*, *113*(29), 8194-8199.
- Li, W. H., Gojobori, T., & Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, *292*(5820), 237-239.
- Ling, G., Miller, D., Nielsen, R., & Stern, A. (2020). A Bayesian framework for inferring the influence of sequence context on point mutations. *Molecular Biology and Evolution*, *37*(3), 893-903.
- Lyons, D. M., & Lauring, A. S. (2017). Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. *Molecular biology and evolution*, *34*(12), 3205-3215.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
- Modiano, G., Battistuzzi, G., & Motulsky, A. G. (1981). Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha-and beta-globin genes: an

evolutionary strategy reducing the rate of mutations with drastic effects?. *Proceedings of the National Academy of Sciences*, *78*(2), 1110-1114.

- Morales, A. C., Rice, A. M., Ho, A. T., Mordstein, C., Mühlhausen, S., Watson, S., ... & Hurst, L. D. (2021). Causes and consequences of purifying selection on SARS-CoV-2. *Genome biology and evolution*, *13*(10), evab196.
- Moriyama, E. N., & Powell, J. R. (1997). Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. *Journal of Molecular Evolution*, *45*, 378-391.
- Muse, S. V., & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution*, *11*(5), 715-724.
- Ngandu, N. K., Scheffler, K., Moore, P., Woodman, Z., Martin, D., & Seoighe, C. (2008). Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. *Virology journal*, *5*, 1-11.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, *60*(359-367), 489-498.
- Purvis, A., & Bromham, L. (1997). Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. *Journal of Molecular Evolution*, *44*, 112-119.
- Rocha, E. P., Touchon, M., & Feil, E. J. (2006). Similar compositional biases are caused by very different mutational effects. *Genome research*, *16*(12), 1537-1547.

- Schroeder, J. W., Randall, J. R., Hirst, W. G., O'Donnell, M. E., & Simmons, L. A. (2017). Mutagenic cost of ribonucleotides in bacterial DNA. *Proceedings of the National Academy of Sciences*, *114*(44), 11733-11738.
- Sen, P., Waris, A., Ray, S. K., & Satapathy, S. S. (2020). A web portal to calculate codon adaptation index (CAI) with organism specific reference set of high expression genes for diverse bacteria species. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019* (pp. 319-325). Springer Singapore.
- Sen, P. (2023). *Computational Analysis of Codon Usage Bias, Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences* (Doctoral dissertation, Tezpur University).
- Sen, P., Aziz, R., Deka, R. C., Feil, E. J., Ray, S. K., & Satapathy, S. S. (2022). Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. *Journal of Molecular Evolution*, *90*(1), 114-123.
- Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S., & Bazykin, G. A. (2012). Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Molecular biology and evolution*, *29*(8), 1943-1955.
- Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular biology and evolution*, *23*(1), 7-9.
- Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, *15*(3), 1281-1295.

- Stoltzfus, A., & Norris, R. W. (2016). On the causes of evolutionary transition: transversion bias. *Molecular biology and evolution*, *33*(3), 595-602.
- Sung, W., Ackerman, M. S., Gout, J. F., Miller, S. F., Williams, E., Foster, P. L., & Lynch, M. (2015). Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. *Molecular biology and evolution*, *32*(7), 1672-1683.
- Tamura, K. (1992). The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. *Molecular biology and evolution*, *9*(5), 814-825.
- Thorpe, H. A., Bayliss, S. C., Hurst, L. D., & Feil, E. J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, *206*(1), 363-376.
- Topal, M. D., & Fresco, J. R. (1976). Complementary base pairing and the origin of substitution mutations. *Nature*, *263*(5575), 285-289.
- Van den Eynden, J., Basu, S., & Larsson, E. (2016). Somatic mutation patterns in hemizygous genomic regions unveil purifying selection during tumor evolution. *PLoS genetics*, *12*(12), e1006506.
- Yang, Z., & Yoder, A. D. (1999). Estimation of the transition/transversion rate bias and species sampling. *Journal of Molecular Evolution*, *48*, 274-283.
- Wolfe, K. H., Sharp, P. M., & Li, W. H. (1989). Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution*, *29*, 208-211.
- Wu, C. I., & Li, W. H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences*, *82*(6), 1741-1745.
- Zhu, Y., Neeman, T., Yap, V. B., & Huttley, G. A. (2017). Statistical methods for identifying sequence motifs affecting point mutations. *Genetics*, *205*(2), 843-856.

- Zou, Z., & Zhang, J. (2021). Are nonsynonymous transversions generally more deleterious than nonsynonymous transitions?. *Molecular biology and evolution*, *38*(1), 181-191.