Higher frequency of transition mutation over transversion mutation in genomes: evidence from the binding energy calculation of base pairs using DFT

Nishita Deka [1,7], Nand Kishor Gour[2], Pradeep Pant[3] Siddhartha Shankar Satapathy[4,6,7]Nasimul Hoda[5], Ramesh Chandra Deka*[2,6,7], Suvendra Kumar Ray* [1,6,7]

[1]*Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, 784028 Assam, India.*

[2]*Department of Chemical Sciences, Tezpur University, Tezpur, 784028 Assam, India.*

[3]*Department of Biotechnology, Bennett University, Greater Noida 201310, Uttar Pradesh*

[4]*Department of Computer Science & Engineering, Tezpur University, Tezpur, 784028 Assam, India.*

[5]*Department of Chemistry, Jamia Millia Islamia, Jamia Nagar,110025, New Dehli*

[6]*Centre for Multidisciplinary Research, Tezpur University, Tezpur, 784028 Assam, India.*

[7]*Centre for Bioinformatics & Computational Biology, Tezpur University, Tezpur, 784028 Assam, India.*

**For Correspondence: suven@tezu.ernet.in*

ramesh@tezu.ernet.in

Abstract

Base substitution mutations such as transition (*ti*) and transversion (*tv*) in organisms are major driving force in molecular evolution. In this study, different possible types of base pairing that can cause *ti* and *tv* were investigated using the density functional theory (DFT) method. The chemical structures of bases as well as base pairs were optimized using B3LYP hybrid functional along with 6-31G(d,p) basis set. We performed single point energy calculation of all optimized species using the same functional but combined with higher diffuse and polarized basis set i.e. 6-311++G(d,p) to get more refined energy of all species. The binding energy of various base pairs was calculated considering basis set superposition error (BSSE) as well as without BSSE. The binding energy of the base pairs leading to *ti* was found to be more stable than that of the base pairs leading to *tv*. This was interesting considering the observations in organisms that *tis* are more frequent than *tvs*. Among the base pairs leading to the same *ti*, G(*keto*): T (*enol)* base pair was found to be more stable than A(*imino*):C(*amino)* base pair. This theoretical study of binding energy of different base pairs using the DFT method has provided additional evidences in support to the biological observations of a higher transition rate than transversion in genomes.

Keywords: Base substitution mutation, Base pairs, Transition, Transversion, DFT, BSSE, Binding energy, Base tautomer, Base conformers.

1. Introduction

Base substitution mutation in DNA is the most common mechanism of genome variation in organisms. These mutations have been classified into two types such as transitions (*ti*) where, a purine (or a pyrimidine) base is substituted by another purine (or pyrimidine) base, and transversions, where a purine base is substituted by a pyrimidine base, or the *vice versa* (Sankoff et al., 1976).

Considering that each base in DNA can be replaced by any of the other three bases, there are a total of twelve possible substitutions, of which eight are transversions and four are transitions (Sankoff et al., 1976). The transitions are C→T, T→C, A→G, and G→A which result from purine-pyrimidine (Pu:Py) base pairing. In transversions, purine-purine (Pu-Pu) or

pyrimidine-pyrimidine (Py-Py) base pairing (Leonard et al.,1994) leading to the following transversions: A→T, A→C, T→A, T→G, G→T, G→C, C→A, and C→G. Although, theoretically transversions should occur more frequently, transition mutations are observed to occur two to three times more frequent than transversions in genomes (Vogel & Kopun, 1977;(Sen et al., 2021); Aziz et al., 2022). This disparity has been mainly attributed to the distortion of the DNA helix geometry caused by Pu-Pu and Py-Py pairs responsible for *tv* compared to Pu-Py pair responsible for *ti* in genomes (Topal & Fresco, 1976). Further, among the different *ti*s, the frequencies of C→T (or G→A) is higher than that of T→C (or A→G) (Kreutzer & Essigmann, 1998). There is also difference among the eight *tv*s: G→T (or C→A) usually occurs at a higher frequency than that of the other *tv*s (Hollstein et al., 1991). Base tautomer, such as amino/imino and keto/enol forms, and base conformers, like anti/syn conformations, play a significant role in forming base pairs that lead to these substitutions (Table 1). Tautomerization primarily contributes to transitions, while the tautomer along with syn conformers facilitate transversions (Topal & Fresco, 1976). Though factors such as DNA geometry, DNA error repair enzymes, damage to DNA bases, RNA secondary structure and codon degeneracy have been implemented towards *ti*s and *tv*s in genomes, there is no report in the literature regarding the base pair stability in DNA with their contribution towards these mutation types. By estimating the binding energies of these base pairs (Table 1), their stability can be determined — the lower the binding energy, the more stable the pairing.

The binding energy of DNA base pairs can be estimated through experimental methods and computational analysis using quantum chemical approaches. The binding energies of the A: T and G:C base pairs have been empirically estimated with values of -13.00 kcal/mol and -21.00 kcal/mol, respectively (Gould & Kollman, 1994; Yanson et al., 1979). With advancements in quantum chemical studies, researchers are now able to estimate the binding energies of base pairs that are difficult to measure experimentally. For example, rare tautomer and base conformers are very short-lived, making it challenging to experimentally determine their stability. Quantum chemical approaches have proven to be promising, as they can mimic experimental results. Several researchers have explored the binding energy between DNA base pairs using computational methods, yielding results similar to experimental data. For example, the binding energy of the A: T base pair has been reported to range from -9.00 to -17.00 kcal/mol (Souri & Mohammadi, 2017; Masoodi et al., 2019), while the binding energy of the G:C base pair has been found to range from -22.00 to -31.00 kcal/mol (Asensio et al., 2003; Acosta-Silva et al., 2010; Thoa & Hue, 2018). Masoodi

et al. (2018) investigated the binding energies of A *imino*: C *amino* and A *amino*: C *imino* base pairs using the DFT method, reporting values of -21.00 kcal/mol and -14.00 kcal/mol, respectively. However, base pairs formed by tautomerization as well as base conformational changes (anti/syn), which contribute to base substitution mutations, have not been extensively studied previously. In this work, we have explored binding energy of different base pairs that have undergone tautomerization, conformational changes, or both. The base pairs we considered (that cause mutation) are based on the hypothesis of Topal and Fresco (1976). Unlike earlier studies, which typically excluded the sugar-phosphate backbone and focused solely on the bases, we included the nucleotide pair, complete with its sugar-phosphate backbone, to more accurately represent the base conformers. We have observed that base pairs leading to tis have lower binding energy than that of the base pairs leading to tvs. This approach has allowed us to provide another explanation for the higher transition-to-transversion *(ti/tv)* ratio in genomes, supported by quantum chemical analysis using the DFT method.

12 substitutions	Base pairs
A→C	A: G
A→G	G: T or A:C
A→T	A: A
C→A	G: A
C→G	G: G
C→T	G: T or A:C
G→A	G: T or A:C
G→C	G: G
G→T	A: G
T→A	A: A
T→C	A:C or G: T
T→G	A: G

Table 1. Base pairs that cause different types of base substitution mutation are shown

2. Materials and Computational Details

All the base structures were drawn using Gauss View 6, starting from nucleotide templates provided by the software and applying the desired modifications. We assumed that the nucleotide templates were accurate and proceeded accordingly. After that, the hydrogen bond is added between the respective pairs. All DFT investigations were executed using the GAUSSIAN 16 software package.(Frisch et al., 2016). Geometry optimizations of single bases and base pairs were conducted using B3LYP functional (Lee et al., 1988) along with 6-31G (d, p) basis set (Muzomwe et al., 2012),(Bouzzine et al., 2005),(Tirado-Rives &

Jorgensen, 2008) B3LYP functional–incorporates Becke's three-parameter hybrid exchange functional and the Lee-Yang–Parr correlation functional. Vibrational frequency calculations are further performed at the same level of theory to get different modes of vibrational frequency as well as verify all the stable minima to have positive frequencies. Single point energy calculation of all optimized species were further performed at the same functional but using higher diffuse and polarized basis set i.e. 6-311++G (d,p) to get more refined energy of all species (Andersson & Uvdal, 2005) ,(Appell et al., 2005) ,(Schnupf et al., 2007). Corrections for basis set superposition error (BSSE) and zero-point vibrational energies (ZPE) for binding energies were taken into account by using the counterpoise (CP) method (Gutowski et al., 1993),(Tao & Pan, 1991),(Novoa et al., 1994)

The binding energy (ΔE) was determined as the difference in energy between the dimer and its isolated monomers. All calculations included counterpoise corrections for BSSE and adjustments for zero-point vibrational energies.

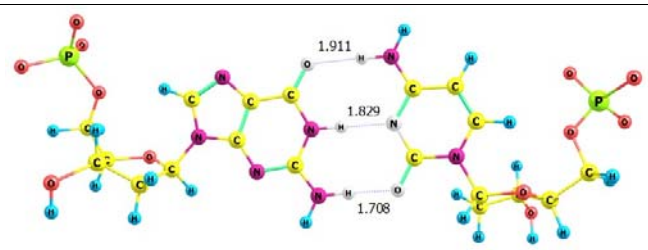
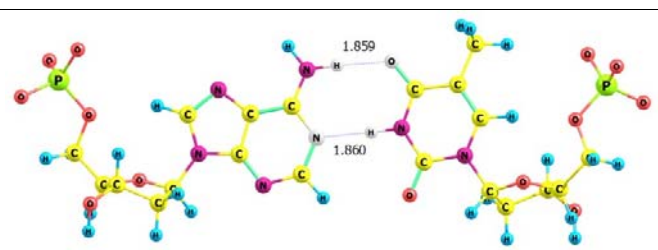
$$B.E. = E_{complex} - (E_{base\,1} + E_{base\,2})$$

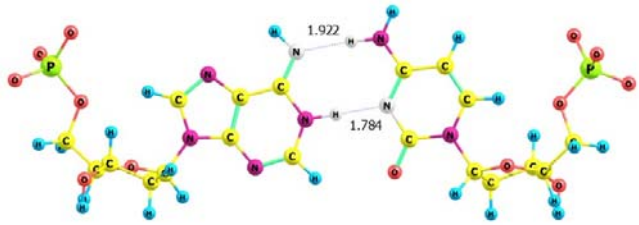
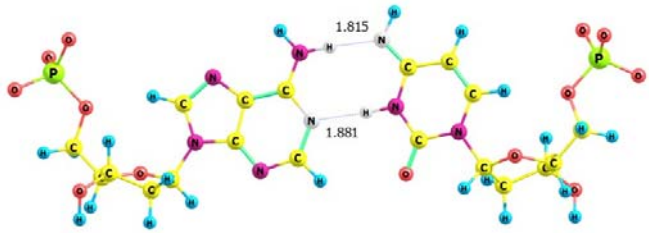
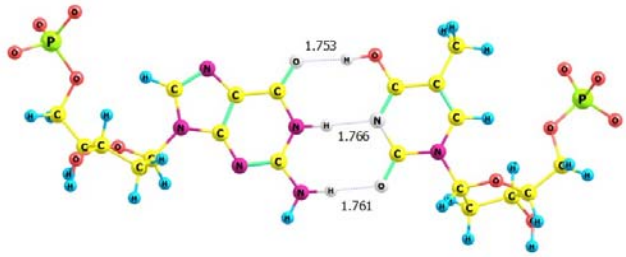
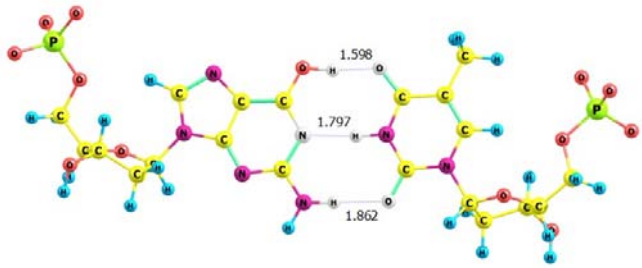
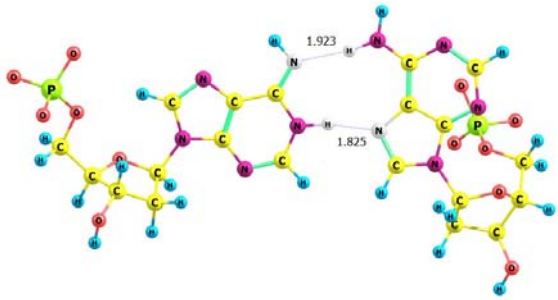
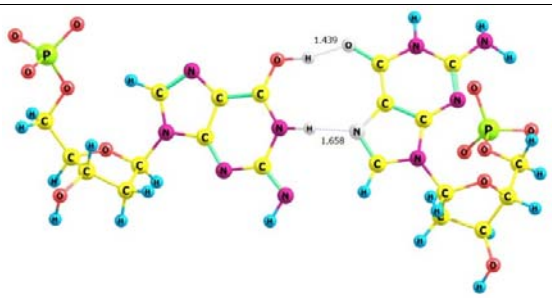
Where E_{base1} and E_{base2} denotes monomer (Single base)

$E_{complex}$ = denotes dimer (base pair)

3. Results

Binding energy of base pairs leading to transition is more stable than the base pairs leading to transversion

(a)A: T	
(b)G:C	

(i) A*: C	
(ii) A: C*	
(iii)**G: T***	
(iv) G*: T	
(v)A* $_{anti}$**: A**$_{syn}$	
(vi) G$_{anti}$**: G**$_{syn}$	

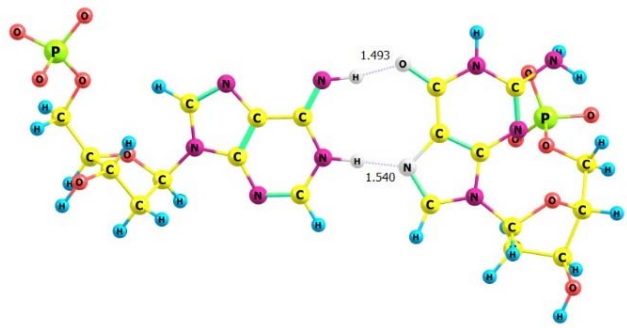
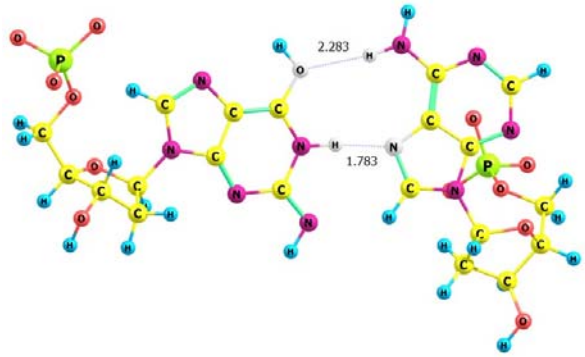
(vii) A*$_{anti}$:G $_{syn}$	
(viii)G$_{anti:}$ A $_{syn}$**	

Figure 1. The optimized A: T and G:C pairs are showing in figure (a) and figure (b) respectively. In figures (i) to (iv) types of base pairs that can lead transition mutations are shown. Whereas, in Figure (v) to (viii) base-pairs that are responsible for transversion mutations are shown. * denotes rare tautomeric form, ** denotes the double tautomeric form.

We analyzed energy of the eight different base pairs that results twelve base substitutions (Table 1). The optimized structures of the eight base pairs along with the A: T and G:C pairs are shown in Figure 1. In the structure, each base pair is held by hydrogen bonds (H-O and H-N). The hydrogen bond distance range was found from 1.4 to 2.2 Å. The coordinates of single bases and base pairs are given in supplementary file 1.

Vibrational frequency calculations are performed to get different modes of vibrational frequency. It was found that all the single bases and base pairs have positive frequency which indicates that all molecule are in ground-state configuration. All the optimized structure corresponds to a true energy minimum on the potential energy surface (PES), meaning that the forces acting on atoms are minimized and the system is in a stable configuration. Vibrational frequencies of single bases and base pairs are given in the supplementary file 2.

Binding energy of each pair was calculated, (all calculation data is given in the supplementary file_3). Binding energy was calculated considering BSSE as well as without

considering BSSE to make the observation stringent. The values of the binding energy for these ten base pairs (including A: T, G:C) are shown in Table 2**.**

Base pairs	B.E (without BSSE)	B.E (with BSSE)	Experimental Value
A: T	-11.08	-10.27	-13.00
G: C	-30.11	-28.98	-21.00
A: C*	-13.17	-12.29	NA
A*: C	-19.91	-19.13	NA
G*: T	-17.35	-16.18	NA
G: T*	-26.98	-25.74	NA
A *: A $_{syn}$	-10.60	-9.77	NA
G **: G $_{syn}$	-17.74	-16.40	NA
A *: G $_{syn}$	-4.46	-3.65	NA
G **: A $_{syn}$	-9.63	-8.75	NA

Table 2: The binding Energy of DNA pairs was calculated using a 6-311++G (d, p) basis set, is shown without considering the BSSE and considering BSSE. All values are written in kcal/mole. Experimental energy value of A: T and G:C are mentioned. * denotes the rare tautomeric form., ** denotes the double tautomeric form.

The binding energies of A: T and G:C pairs were found to be -10.27 kcal/mol and -28.98 kcal/mol, respectively. The higher stability of G:C pair than A: T is expected considering three hydrogen bonds between G and C in comparison to two hydrogen bonds between A and T. The energy calculations of these base pairs are within the range that has been reported earlier (Souri & Mohammadi, 2017; Masoodi et al., 2019, Asensio et al., 2003; Acosta-Silva et al., 2010; Thoa & Hue, 2018). The base pairs that can lead to transitions are A*(imino*)**:** C (*amino*) (Fig. 1. (i), A(*amino*)**:** C(*imino*) (Fig. 1. (ii)), G(*keto*): T*(enol*) (Fig. 1. (iii)), G(*enol*)**:** T (*keto*) (Fig. 1. (iv)), whose binding energies were -19.13 kcal/mol, -12.29 kcal/mol, -25.74 kcal/mol and -16.18 kcal/mol, respectively (Table 2). A noticeable energy difference observed between the A(*imino*)**:**C(*amino*) and A(*amino*)**:**C(*imino*) pairs that results in G:C →A: T and A: T →G:C, respectively. Similarly, there is an energy difference between G(*keto*)**:** T*(enol*) and G(*enol*)**:** T (*keto*) base pairs that result in G:C →A: T and A: T →G:C transitions,

respectively. So according to the binding energy, transition mutations favoring G:C → A: T is more preferred in comparison to A: T → G: C. This is interesting because by doing genome sequence analysis several researchers have observed base substitution mutation in organisms is biased towards AT (Hershberg & Petrov, 2010),(Beura et al., 2023),(Sen et al., 2021) .

Different base pairs such as A (*imino*)$_{anti}$: A$_{syn}$ (Fig. 1(v)), G (*enol, imino*) $_{anti}$: G$_{syn}$ (Fig. 1(vi)), A (*imino*)$_{anti}$: G$_{syn}$ (Fig. 1(vii)), and G (*enol, imino*) $_{anti}$: A$_{syn}$ (Fig. 1(viii)) (Topal & Fresco, 1976) can cause transversion in a genome. The binding energies of these four base pairs were estimated as -9.77 kcal/mol, -16.40 kcal/mol, -3.65 kcal/mol, and -8.75 kcal/mol (Table 2), respectively. It indicated a wide variation of their energy values. There is no available data in the literature regarding the energy values of these bases pairs that could be used as our reference to compare. However, the energy values indicate a greater magnitude of difference across different transversions occurring in organisms.

We compared the stability of base pairs leading to transition with that of base pairs leading to transversion. All four pairs resulting into transition have the energy below -10.00 Kcal/mol, whereas in the case of transversion, only one base pair has energy below -10.00 kcal/mol. This observation clearly indicates that, the base pairs that result to transition mutation are more stable than that of the base pairs that result to transversions. This observation supports the high *ti* frequency than *tv* in genomes. So far, we understand this is the first DFT comparative study of base pairs supporting the higher frequency of transition than transversion in genomes.

4. Discussion

Base substitution mutations in DNA are of two types such as transition and transversion. Transition results from four different base-pairs and transversion results from eight different base-pairs. However, in genomes, transition is observed to be more frequent than transversion that has been attributed to DNA geometry, RNA secondary structures, DNA repair, and codon degeneracy. However, in terms of energy of base pairs leading to these different mutations have not been reported earlier. In this study, we have calculated the stability of different possible base pairs that have the potential to cause transition as well as transversion mutations. The observations indicate that base pairs leading to *ti* are more stable than that of the base pairs leading to *tv*. So, in the present work using the DFT method we

have provided another evidence in terms of binding energy in favor of the higher frequency of *ti* over *tv*. As per our knowledge, this is the first time explanation by DFT calculations regarding higher frequency of *ti* than *tv* in a genome. Though, the hypothesis by Topal and Fresco (1976) was given five decades earlier, the earlier computer programs might have been unable to run such small energy calculations, which is possible now by quantum mechanics.

Base pairs that involve tautomer such as A(*imino*):C or A:C(*imino*) and G(*enol*): T or G: T(*enol*) can result into transition mutations. Determining the stability of these base pairs experimentally, is challenging due to the short-lived nature of tautomer. However, computational studies enable accurate stability estimation by simulating experimental conditions. For transversion base pairs, both tautomer and base conformer are involved, further complicating experimental stability assessments. Quantum chemical studies, however, provide high-accuracy stability predictions.

Among the four different transitions, the C→T as well as G→A are more frequent than the other two (Kreutzer & Essigmann, 1998). Though our study on base pair stability is in support of this, there are other reasons such as frequent cytosine deamination might be attributed to the high C→T frequency (Coulondre et al., 1978). Similarly, if we observed the base pair stability in the case of *tv*, the G→T transversion was found to be less frequent. But in nature, the G→T frequency is high because of the guanine oxidation resulting in 8-oxo-guanine which causes G→T transversion(Andrés et al., 2023). Therefore, apart from the stability of base pairs, the frequency of different forms of the bases as well as the damaged bases are important attributes determining the frequency of different base substitutions.

It is pertinent to note that in this study, we analyzed energy of single base pairs only and it cannot be explained in DNA in terms of geometry, but future work could involve oligonucleotides to assess the overall stability of DNA strands with point mutations, providing insights into mutation-induced changes in DNA stability and double-helix geometry.

Authorship contribution statement

Nishita Deka: Conceptualization, Methodology, Writing- Original draft preparation, Software, Visualization, Investigation. **Nand Kishor Gour**: Methodology, Data curation, Writing- Original draft preparation, Software, Validation. **Pradeep Pant:** Validation, Writing- Reviewing and Editing **Siddhartha Shankar Satapathy, Nasimul Hoda:** Reviewing and Editing**, Ramesh Chandra Deka, Suvendra Kumar Ray**: Supervision, Conceptualization, Methodology Software, Validation, Writing- Reviewing and Editing

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgement

ND is thankful to Dr. Dikshita Dowerah for her Guidance in initial studies. ND also acknowledges DBT, GoI-funded Centre for Bioinformatics and Computational Biology, Tezpur University (BT/PR40253/BTIS/137/52/2022) for the fellowship of Scientific Administrative Assistant. ND and SKR is also thankful to the computation facility of Centre for Multidisciplinary Research (CMDR) of Tezpur University and the NNP grant awarded to SKR.

References

Andersson, M. P., & Uvdal, P. (2005). New Scale Factors for Harmonic Vibrational Frequencies Using the B3LYP Density Functional Method with the Triple-ζ Basis Set 6-311+G(d,p). *The Journal of Physical Chemistry A*, *109*(12), 2937–2941. https://doi.org/10.1021/jp045733a

Andrés, C. M. C., Lastra, J. M. P. de la, Juan, C. A., Plou, F. J., & Pérez-Lebeña, E. (2023). Chemical Insights into Oxidative and Nitrative Modifications of DNA. *International Journal of Molecular Sciences*, *24*(20), 15240. https://doi.org/10.3390/ijms242015240

Appell, M., Willett, J. L., & Momany, F. A. (2005). DFT study of α- and β-d-mannopyranose at the B3LYP/6-311++G** level. *Carbohydrate Research*, *340*(3), 459–468. https://doi.org/10.1016/j.carres.2004.12.010

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), 1–7. https://doi.org/10.1007/s12041-023-01418-w

Bouzzine, S. M., Bouzakraoui, S., Bouachrine, M., & Hamidi, M. (2005). Density functional theory (B3LYP/6-31G*) study of oligothiophenes in their aromatic and polaronic states. *Journal of Molecular Structure: THEOCHEM*, *726*(1–3), 271–276.

Coulondre, C., Miller, J. H., Farabaugh, P. J., & Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, *274*(5673), Article 5673. https://doi.org/10.1038/274775a0

Frisch, M. ea, Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, Ma., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., & Nakatsuji, H. (2016). *Gaussian 16*. Gaussian, Inc. Wallingford, CT.

Gutowski, M., van Duijneveldt-van de Rijdt, J. G. C. M., van Lenthe, J. H., & van Duijneveldt, F. B. (1993). Accuracy of the Boys and Bernardi function counterpoise method. *The Journal of Chemical Physics*, *98*(6), 4728–4737. https://doi.org/10.1063/1.465106

Hershberg, R., & Petrov, D. A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics*, *6*(9), e1001115. https://doi.org/10.1371/journal.pgen.1001115

Hollstein, M., Sidransky, D., Vogelstein, B., & Harris, C. C. (1991). P53 Mutations in Human Cancers. *Science*, *253*(5015), 49–53. https://doi.org/10.1126/science.1905840

Kreutzer, D. A., & Essigmann, J. M. (1998). Oxidized, deaminated cytosines are a source of C → T transitions *in vivo*. *Proceedings of the National Academy of Sciences*, *95*(7), 3578–3582. https://doi.org/10.1073/pnas.95.7.3578

Lee, C., Yang, W., & Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, *37*(2), 785–789. https://doi.org/10.1103/PhysRevB.37.785

Muzomwe, M., Maes, G., & Kasende, O. E. (2012). Theoretical DFT(B3LYP)/6-31+G(d) study on the prediction of the preferred interaction site of 3-methyl-4-pyrimidone with different proton donors. *Natural Science*, *4*(5), Article 5. https://doi.org/10.4236/ns.2012.45041

Novoa, J. J., Planas, M., & Whangbo, M.-H. (1994). A numerical evaluation of the counterpoise method on hydrogen bond complexes using near complete basis sets. *Chemical Physics Letters*, *225*(1), 240–246. https://doi.org/10.1016/0009-2614(94)00646-6

Sankoff, D., Cedergren, R. J., & Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *Journal of Molecular Evolution*, *7*(2), 133–149. https://doi.org/10.1007/BF01732471

Schnupf, U., Willett, J. L., Bosma, W. B., & Momany, F. A. (2007). DFT study of α- and β-d-allopyranose at the B3LYP/6-311++G∗∗ level of theory. *Carbohydrate Research*, *342*(2), 196–216. https://doi.org/10.1016/j.carres.2006.12.006

Sen, P., Aziz, R., Das, S., Namsa, N. D., Deka, R. C., Feil, E. J., Ray, S. K., & Satapathy, S. S. (2021). *Polymorphism at four-fold degenerate site, but not at the intergenic regions, explains nucleotide compositional strand asymmetry in bacteria* (p. 2021.10.07.463495). bioRxiv. https://doi.org/10.1101/2021.10.07.463495

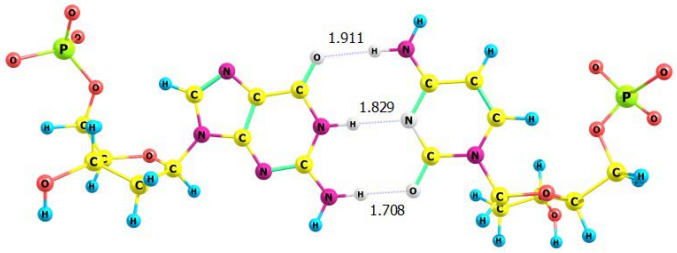
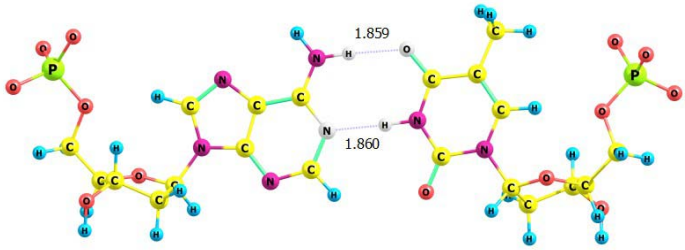
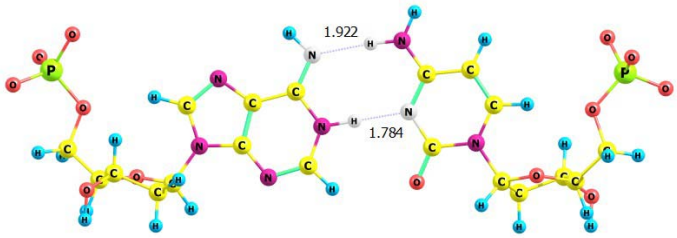
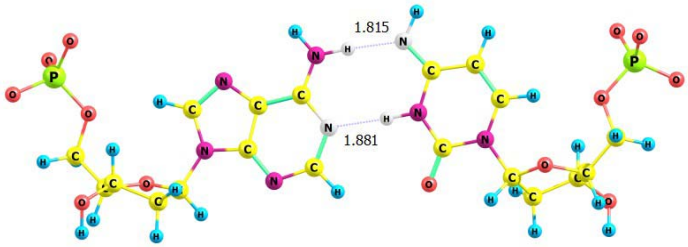
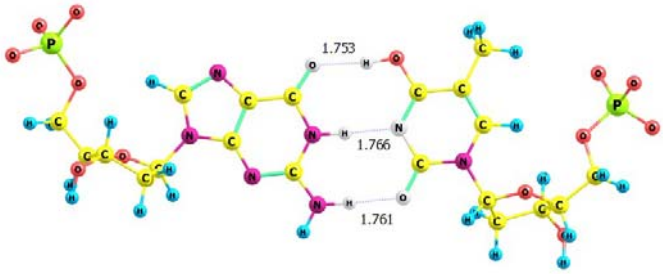
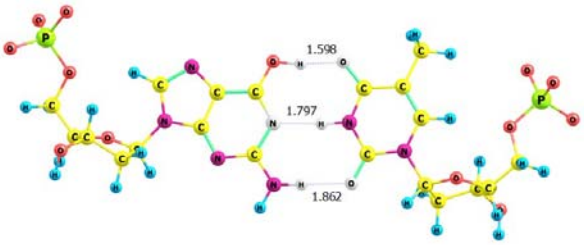
Tao, F. M., & Pan, Y. K. (1991). Validity of the function counterpoise method and ab initio calculations of van der Waals interaction energy. *The Journal of Physical Chemistry*, *95*(9), 3582–3588. https://doi.org/10.1021/j100162a028

Tirado-Rives, J., & Jorgensen, W. L. (2008). Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules. *Journal of Chemical Theory and Computation*, *4*(2), 297–306. https://doi.org/10.1021/ct700248k

Topal, M. D., & Fresco, J. R. (1976). Complementary base pairing and the origin of substitution mutations. *Nature*, *263*(5575), 285–289. https://doi.org/10.1038/263285a0

Figure Legends

Figure 1. The optimized A: T and G:C pairs are showing in figure (a) and figure (b) respectively. In figures (i) to (iv) types of base pairs that can lead transition mutations are shown. Whereas, in Figure (v) to (viii) base-pairs that are responsible for transversion mutations are shown. * denotes rare tautomeric form, ** denotes the double tautomeric form.

(a)A: T	
(b)G:C	
(i) A*: C	
(ii) A: C*	
(iii)**G: T***	
(iv) G*: T	

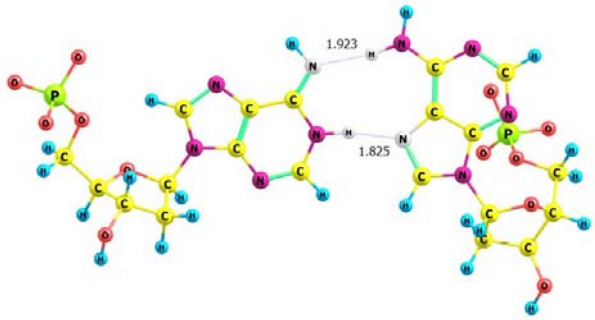
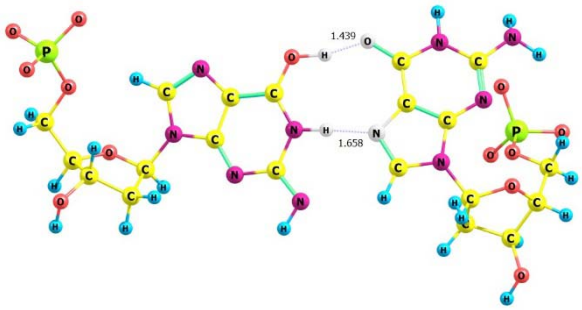
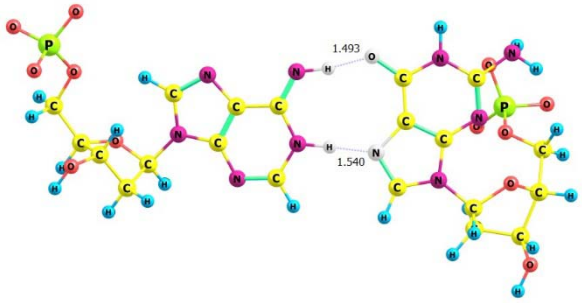
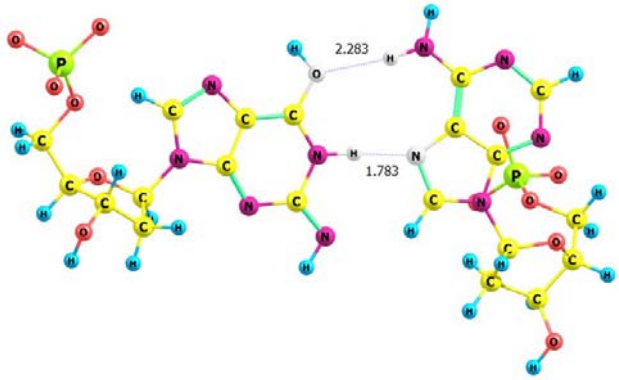
(v)A* $_{anti}$**: A**$_{syn}$	
(vi) G$_{anti}$**: G**$_{syn}$	
(vii) A*$_{anti}$**:G** $_{syn}$	
(viii)G$_{anti:}$ **A** $_{syn}$	

Figure 1: The optimized A: T and G:C pairs are showing in figure (a) and figure (b) respectively. In figures (i) to (iv) types of base pairs that can lead transition mutations are shown. Whereas, in Figure (v) to (viii) base-pairs that are responsible for transversion mutations are shown. (* above the bases denotes rare tautomeric form, ** denotes the double tautomeric form.)