A Numerical Representation and Classification of Codons to Investigate Codon Alternation Patterns during Genetic Mutations on Disease Pathogenesis

Antara Sengupta[a,*], Subhadip Chakraborty[b], Pabitra Pal Choudhury[d], Swarup Roy[c,*], Jayanta Kumar Das[d,1], Ditipriya Mallick[e], Siddhartha S Jana[e]

[a]*Department of Computer Science and Engineering, University of Calcutta, West Bengal, India*
[b]*Department of Botany, Nabadwip Vidyasagar College, Nabadwip, India*
[c]*Department of Computer Applications,Sikkim University, Gangtok, Sikkim, India*
[d]*Applied Statistical Unit, Indian Statistical Institute, Kolkata, India*
[e]*School of Biological Sciences, Indian Association for the Cultivation of Science, Kolkata, India*

Abstract

Alteration of amino acid is possible due to mutation in codons that could have potential reasons to occur disease. Effective mutation analysis can help to predict the fate of the diseased individual which can be validated later by in-vitro experiments. It may also help an individual who is asymptomatic but having a particular genetic change for early detection and diagnosis during any terminal diseases. We try to investigate the codon alteration patterns and its impact during mutation for the genes known to be responsible for a particular disease. For our current study, we consider two neurodegenerative diseases, Parkinson's disease and Glaucoma. We use numerical representation of four nucleotides based on the number of hydrogen bonds in their chemical structures and make classification of 64 codons as well as amino acids into three different classes (Strong, Weak and Transition). The entire analysis has been carried out based on these classifications. Our analysis reveals that the strong class codons are highly mutated followed by weak and

*Corresponding author (Email: antara.sngpt@gmail.com, sroy01@cus.ac.in)

[1]Present address: Department of Pediatrics, Johns Hopkins University School of Medicine, Maryland, USA

transition class. We observe that most of the mutations occur in the first or second positions in the codon rather than the third. While looking into the chemical properties of amino acid, we observe that amino acids belong to the aliphatic group are affected most during missense mutations. Our investigation further emphasises that in most of the cases the change in the determinative degree of codon due to mutation is directly proportional to the physical density property. Further calculation of determinative degree of 20 amino acids and getting determinative degree of any arbitrary set of amino acid sequences from them may help biologists to understand the evolutionary information of amino acid occurrence frequencies of in proteins. In addition, our scheme gives a more microscopic and alternative representation of the existing codon table that helps in deciphering interesting codon alteration patterns during mutations in disease pathogenesis.

Keywords: Determinative degree, Dual Nucleotide, Codon, Mutation, Physicochemical property, Neurodegenerative disease.

---

1. Introduction

Genes are the functional units of heredity [1]. It is mainly responsible for the structural and functional changes and for the variation in organisms which could be good or bad. DNA (Deoxyribose Nucleic Acid) sequences build the genes of organisms which in turn encode for particular protein using codon. Any fluctuation in this sequence (codons), for example, mishaps during DNA transcription, might lead to a change in the genetic code which alter the protein synthesis. This change is called mutation. Mutations occur in two ways: i) a base substitution, in which one base is substituted for another; ii) an insertion or deletion, in which a base is either incorrectly inserted or deleted from a codon. Base substitutions can have a variety of effects [2]. The silent mutation is an example of a base substitution, where the change in nucleotide base has no outward effect. For example, $TTT \Rightarrow TTC$ change ultimately has no effect on protein as both the codons code for the amino acid Phenylalanine. A missense mutation refers to a base substitution where alteration in any nucleotide bases alters the amino acid [3, 4]. For instance, $ACA \Rightarrow AAA$ alteration causes the alteration in coded amino acid from Threonine to Lysine. A nonsense mutation refers to a base substitution in which the changed nucleotide transforms the codon into a stop codon. Such a change leads to a premature termination of translation.

$TAC \Rightarrow TAA$ change leads to Tyrosine to stop codon. If a protein with 400 amino acids having this nonsense mutation on the amino acid number say 300 then the protein will terminate with total amino acid of 300 (premature termination due to stop codon). This will always cause the disease phenotype as the full protein is not being translated due to premature termination. This is very interesting for missense mutation that a mutated protein which differs from its wild type counter parts by only one amino acid (due to mutation) and causing the disease phenotype [5].

The alteration of amino acid due to the alternation in codon during disease may not be an arbitrary phenomenon [6]. They may follow certain pattern of alternation. It is an interesting issue to investigate any hidden or priorly unknown alternation patterns, if any, during mutation. This may help geneticist to understand better the mechanism of mutation in disease phenotype. We feel that to elucidate impact of codon alteration or pattern of alteration, positional impact of codon is also equally important. Hence, it is essential to make suitable numerical representation of the biological facts, features or characteristics for the ease of quantitative and computational analysis.

Plethora of contributions are dedicated towards characterizing genes through the light of numerical representations [7, 8, 9, 10]. People have tried in different ways to construct mathematical model for similarity analysis of nucleotide sequences [11, 10, 12]. Rumer in his work has introduced the existence of a symmetry in the genetic code, which is also known as Rumer's transformation or Amino-Keto transformation [9]. The transformation method applied to codon changes the degeneracy of the codon from its original one. Gonzalez et al have constructed mathematical models of the genetic code [13, 14, 15, 16]. According to them Rumer's symmetry could have originated in an ancestral version of the genetic code [17]. Numeric representation of each type of multiplet signifies the number of codons code for that particular amino acid. However, above numerical representation give less importance in hidden abilities (if any) of nucleotides, their positional impact in codon to participate in mutations.

In this present work as a first step of our analysis, we propose a numerical scheme for representing codons in a more effective and meaningful manner, giving due importance to the nucleotides and their positions in a codon. At first we calculate a determinative degree of each codon followed by classifying 64 codons into three different classes (weak, transitional, strong) based on their strength in codon degeneracy. We next calculate degree of each

amino acids based on above determinative degree each constituent codons and classify 20 amino acids into above classes. We classify 64 amino acids into three classes based on our proposed determinative degree score. We proposed for determinative degree of 64 codons followed by degree of 20 amino acids and determinative degree of any arbitrary amino acid sequence. The overall idea of the numeric representation and codons and amino acids classification is to measure the strengths of nucleotides, role of the third nucleotide and their impact during mutations causing fatal diseases. We use our scheme to understand the pattern of codon alteration in two neurodegenerative diseases namely, Parkinson and Glaucoma[18] and their effects in the physical and chemical properties at DNA primary sequential level as well as secondary structural level. Proximity between set of any arbitrary amino acid sequences can be measured based on determinative degrees they have.

2. Methodology

The strengths of nucleotides in DNA sequential level have deep impact in protein formation and thus alteration may cause genetic diseases. It is therefore important to analyse the impact of positional alteration of nucleotides in codon during mutation. Due to degeneracy factor a codon may codes for more than one amino acids, leads to multiple structure of amino acids [7]. Rumer [9] in his seminal work tried to explain the degeneracy with the help of first two constituent nucleotides (diletters) of a codon, termed as *root* and classify the roots into three different classes, namely *Strong*, *Transitional* and *Weak*. Due to multiplate structure of codons, codons code for the same amino acids can be separated from root. For example, if we consider xyz as codon, then root xy can be separated from z. Thus, with four nucleotide bases, we get $4^2 = 16$ possible roots, which he represented using a rhombic structure. Rumer arranged four nucleotides (Adenine (A), Thymine (T), Guanine (G), Cytosine C)) in chronological order according to their strength which is **C, G, T, A**. Rumer's representation is good in analysing impact of positional alteration. However, Rumer's model is limited only with dual nucleotides. Rumer's scheme is incomplete in nature in the sense it fails to classify all the available codons.

2.1. Numerical representation of codon

We extend the existing dual nucleotide representation towards triplet or whole codon. We shift the concept from *root* to $root + 1$, where root holds

the first two positions of a codon. The first two positions of a codon having greater importance to code any amino acid. Hence, due to mutation when a codon get changed, there is a possibility of its root to be shifted from one class to the other.

Dual Nucleotide	Amino Acid	Average Molecular Weight	Positional Specification in the Rhombic Structure
CC	P	115.1	1
GC	A	89.1	1
CG	R	174.2	2
GG	G	75.1	1
TC	S	105.1	2
CT	L	131.2	3
GT	V	117.1	1
AC	T	139.65	2
CA	H,Q	150.7	3
TG	C,W	162.7	4
AG	S,R	139.65	1
GA	D,E	140.1	2
TT	F,L	148.2	3
AT	I,M	140.2	1
TA	Y	181.2	2
AA	N,K	139.15	1

We have represented each codon using three digits number as shown in . Each digit in three digits number signifies different properties of a codon. The first position indicates the additive degree of the root or group number. The first step towards for any representation scheme is to assign numerical weights to four nucleotides before representing codon using any number. In support of Rumer's ordering of four nucleotides [19] assign a determinative degree for every nucleotides (an abstract characteristic of nucleotides) where $d_C = 4$, $d_G = 3$, $d_T = 2$, and $d_A = 1$. Thus, each dual nucleotides can have range of numeric values from 2 to 8 only. For example, root **CC** scores $4 + 4 = 8$; **AA** scores $1 + 1 = 2$. If we consider the Rumer's rhombic arrangements of dual nucleotides we can have seven classes of dual nucleotides based on their additive scores. The second digit represents the position of the root in a particular row in the rhombic structure. We use positional specification at second position of codon to provide unique identity to each codon, else additive degree of more than one codon will be same. In order to represent a codon uniquely, we introduce a positional specification of each doublet or root in the rhombic organization of roots.To do so, we have recorded the average score of molecular weights of the amino acids they code(Table **??**. The positions of dual nucleotides at each row are numbered accordingly. The last digit is the determinative degree of the third nucleotide. The format of

numerical representation of codon table is shown in Fig 2.

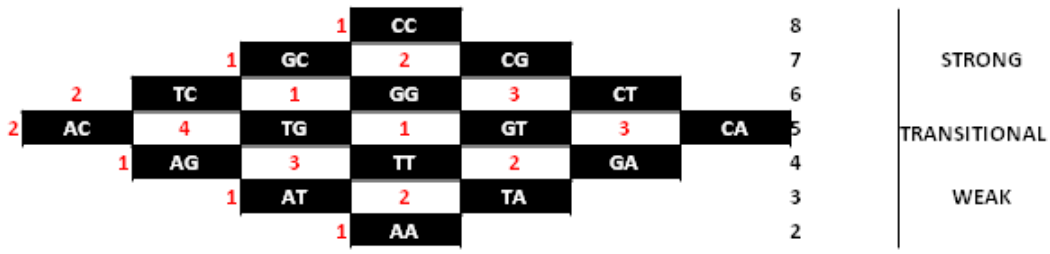


Figure 1: Rhombic structure of roots. Order of three dual nucleotide properties (Bottom to top): Week, Transitional and Strong. The positional specifications of each dual nucleotide are recorded according to the average score of molecular weights of the amino acids they code).

As a example, for codon CCA, nucleotides 4(C) at first and second positions of codon will form a dual nucleotide CC (4+4=8) at first step. Thus getting the numeric value 8 at first position of the determinative degree, it will go to next step to get the numeric value at second position of it, which is 1 as per rhombic structure specified in Figure 1. So now at the end of step 2, the value is 81. At third step it will select the numeric value 1 for the nucleotide at third position, which is A. So now the determinative degree of the codon CCA is 811. It codes for the amino acid P(Proline). As according to the classification of dual nucleotide the CC comes under group G8 and class of Strong(S).We termed this three digit representation of codon as degree of codon and formally can be given as follows.

Definition 2.1 (Degree of Codon)**.** *Given a codon C with constituent nucleotides say, $\{N_1, N_2, N_3\}$, where $N_i \in N$, the degree of C ($\delta(C)$) is the concatenation (.) of three digits.*

$$\delta(C) = D_1.D_2.D_3, \tag{1}$$

where, D_1= determinative degree of dual nucleotide or additive degree of nucleotides, D_2=Column number in rhombic structure and D_3 is the determinative degree of third nucleotides as discussed above.

2.2. Classification of amino acids

The four nucleotides possessing different bonding strengths based on the number of hydrogen bonds in their chemical structures. According to that

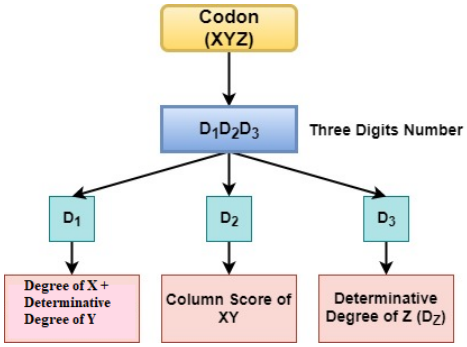


Figure 2: Scheme for three digits numerical representation of codons based on dual nucleotide scores and degenerative degree of third nucleotide.

Table 1: Numeric representation of genetic code table based on determinative degree of 64 codon targeted to twenty amino acids (AA).

DD OF CODON	AA	DD OF CODON	AA	DD OF CODON	AA	DD OF CODON	AA
432	F	622	S	322	Y	542	C
434	F	624	S	324	Y	544	C
431	L	621	S	321	STOP	541	STOP
433	L	623	S	322	STOP	543	W
632	L	812	P	532	H	822	R
634	L	814	P	534	H	824	R
631	L	811	P	531	Q	821	R
633	L	813	P	533	Q	823	R
312	I	522	T	212	N	412	S
314	I	524	T	214	N	414	S
311	I	521	T	211	K	411	R
313	M	523	T	213	K	413	R
512	V	712	A	422	D	612	G
514	V	714	A	424	D	614	G
511	V	711	A	421	E	611	G
513	V	713	A	423	E	613	G

C and **G** can be considered as strong bases whereas rest two i.e. **T** and **A** is weak bases. Again, within a strong and weak base pairs they may have comparative strength within them such that pyrimidine bases, **C** and **T** are stronger than purine bases, **G** and **A**, respectively [20]. We classify all 64 codons into three classes namely Weak, Strong and Transitional according to the additive score of the root as shown in Figure 1. Thus the classification is done based on the additive score of root of each codon.

Accordingly, the codons code for the amino acids are classified too. We find six amino acids (P, A, R, S, G, L) as strong, five amino acids (T, C, W, H, Q) as transitional and rest (F, L, D, E, I, M, Y, N, K, S, R) as weak.

The maximum number of amino acids i.e. eleven (11) are fall under weak class. Unlike, previous attempts we are able to classify all the codons into three different classes. While classifying, we observe an interesting fact that Arginine (R), Serine (S) and Leucine (L) shares two different classes, Weak and Strong. All three amino acids are coded by six (06) different codons, out of which four falls in strong group and rest of two in weak group.

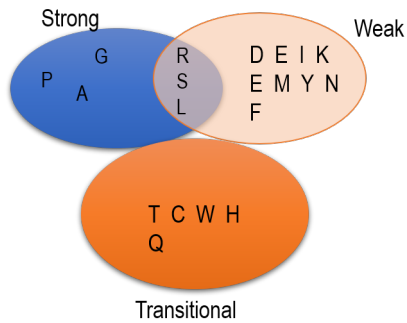


Figure 3: Distribution of 20 amino acids in three different overlapping classes.

2.3. Determinative degree of amino acid

Due to codon degeneracy, it is difficult to represent an amino acid with single numerical value, which is important for subsequent numerical analysis of DNA sequence. To handle the issue we calculate an average score of all the constituent codons for an amino acids, termed as determinative degree. Let $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_n\}$ be the set of twenty ($n = 20$) amino acids. To understand the impact of codons targeted to the amino acid, we calculate the degree (or strength) of amino acid by considering the average degree of all codons that code for a target amino acid.

Definition 2.2 (Degree of Amino Acid)**.** *Let $\mathcal{A}_k$ be an amino acid where $\mathcal{A}_k \in \mathcal{A}$ and $\mathcal{A}_k = \{C_1, C_2, \cdots C_n\}$ such that C_i is the constituent codon codes for $\mathcal{A}_k$. The degree of amino acid $\mathcal{A}_k$ is denoted by $\delta'(\mathcal{A}_k)$ and can be calculated as follows:*

$$\delta'(\mathcal{A}_k) = \frac{\sum_{i=1}^{n} \delta(C_i)}{|\mathcal{A}_k|} \tag{2}$$

We report degree of all 20 amino acids calculated using Equation 2 in Table 2. Interestingly, it can be observed that the amino acids which come

under strong having degree between the range of 612.5 to 812.5. In this way the range of degree scores of transitional group of amino acids lies between 512.5 to 543 and of weak group of it is 212 to 433. It is to be noted that as Ser, Arg and Leu these three amino acids individually are encoded by 6 codons, so among 6 codons, 4 codons come from a codon family, whereas, remaining two codons are codon pairs. According to our classification rule, in all cases 4 codons come from strong root of class and remaining 2 codons come from weak root class. Hence, these three amino acids each has two degrees. Thus with support to the existing codon table our method numerically able to distinguish them, to give more microscopic view.

Table 2: Degree of twenty amino acids in three groups: Strong, Transitional and Weak.

Amino Acid (A_i)	Class	δ'_{A_i}	Amino Acid (A_i)	Class	δ'_{A_i}
Lys		212.00	Ser		413.00
Asn		213.00	Glu		422.00
Tyr		313.00	Asp		423.00
Ile		322.33	Leu		432.00
Met	Weak Root	323.00	Phe	Weak Root	433.00
Arg		412.00			
Val		513.50	Leu		612.50
Thr		522.50	Gly		632.25
Gln		532.00	Ser		622.50
His		533.00	Ala		712.50
Cys	Transitional Root	543.00	Arg	Strong Root	822.50
Trp		543.00	Pro		812.50

2.4. Determinative degree of amino acid sequence

Information about a protein's function and its evolutionary history can be obtained from it's primary structure or amino acid sequence. So it is essential to determine the protein's amino acid sequence [21]. The frequency of occurrence of amino acids constituting an amino acid sequence varies from gene to gene and is well-conserved from species to species [22, 23]. The composition of protein works in such a way that the impact of mutations on protein structure can be minimized. Thus, amino acid composition of proteins reduces deleterious impact of mutations [24]. Moreover, amino acids have different propensities for forming alpha helices, beta sheets, and beta turns [21]. Hence, to understand the impact of occurrence of amino acids in an amino acid sequence on evolution, it is necessary to calculate the degree (or strength) of amino acid sequence by considering the average degree of all amino acids that constituents the amino acid sequence.

Definition 2.3 (Degree of amino acid sequence)**.** *Let $\mathcal{S} = A_1, A_2, ...A_m$ be an arbitrary amino acid sequence of length m, where $\mathcal{A}_k$ be any amino acid, i.e. $A_k \in \mathcal{A}$. The degree of amino acid sequence is denoted by $\delta'(\mathcal{S})$ and can be calculated as follows:*

$$\delta'(\mathcal{S}) = \frac{\sum_{i=1}^{m} \delta(\mathcal{A}_k)}{|\mathcal{A}_m|} \tag{3}$$

3. Results and Discussions

Our prime objective is to analyse codon alteration pattern in disease phenotype. For the ease computational analysis we represent codon numerically. With the help of our scheme we next analyse few mutational DNA and compare their alteration pattern in comparison to non disease phenotypes. We consider four different types of genetic disease for analysis. We perform few more experiments on the mutational data to understand various interesting facts and patterns usually follow during mutation.

3.1. Disease dataset used

We consider two types of genetic diseases namely, Parkinson's Disease and Glaucoma for experimentation. They are Parkinson's Disease and Gloucoma. Parkinson's disease is a progressive disease of the nervous system, where PINK1, PARKIN and DJ1 are three main pathogenic genes [25, 26]. Glaucoma is a condition of increased ocular pressure within the eye, causing gradual loss of sight [27]. MYOC and CYP1B1 are two significant genes responsible for occurring this disease. The wild type sequences of PARKIN (Accession No: AB009973.1), PINK1(Accession No: AB053323.1), DJ1 (Accession No: NM_007262.4), CYP1B1 (Accession No: NM_000104.3), MYOC (Accession No: NM_000261.1) are collected from NCBI GeneBank. We took a group of mutations reported in these genetic diseases collected from various online sources[2],[3],[4], [5], [6].

We consider only the point mutations (missense mutations, nonsense mutations). These are selected in an unbiased way based on the published

[2]http://www.molgen.vib-ua.be/PDMutDB/
[3]http://www.myocilin.com/variants.php
[4]https://databases.lovd.nl/shared/genes/CYP1B1
[5]https://www.omim.org/
[6]https://www.ncbi.nlm.nih.gov/pubmed/

Table 3: Mutation datasets (Missense sample (MS) and nonsense sample (NS)) of neurodegenerative disease types and seven genes used for experimentation.

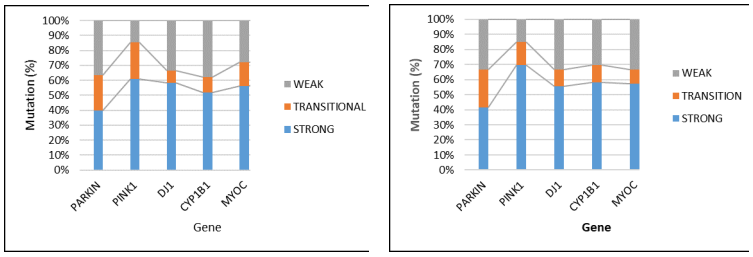
Disease Type	Disease Name	Gene Name	#MS	#NS
Neurodegenerative	Parkinson's Disease	PINK1	41	3
		PARKIN	54	3
		DJ1	12	
	Glaucoma	MYOC	25	5
		CYP1B1	50	

reports for each mutation. More number the mutation reported, more strong the mutation to cause the disease. We considered at least three published reports for each mutation. However, in case of PINK1 and DJ1, we found very less reported evidences of mutation.

3.2. Codon class specific quantification (%) of mutation

Here in this section it is aimed to find the general tendency of mutation taking place in codon classes, whether majority of mutations take place in codons of strong class or weak class or transitional class. The result has been shown in Figure 4. It can be observed from the figure that mutated samples of all types of diseases have a common tendency of mutations in the strong class of codons, i.e. in G8, G7 and G6 . In the case of PINK1 after strong class of codons (60.98%), the mutation has been taken place in transitional class(24.39%). In remaining six genes along with strong class of codons, weak class of codons have remarkable participation.

In addition to that, we also checked the position of mutations in the context of protein domain (active or non active) that may be altered due to mutations. The active domains list is reported in Supplementary Table S1. Interestingly, we observe that mutations mostly present in functional/active region. The overall observation also depicts that the strong class of codons class of codons are affected the most (Figure 4). Moreover, we look into three amino acids Arg, Ser, and Leu, where each of them are coded by six codons, four of them are from strong class and the remaining two codons are from week class. While existing genetic code table only speaks about the name of the mutated amino acids,our rule of classification, we can give more microscopic view. We observe that the maximum mutated codons belong to the strong class except for genes PARKIN, where mutated codons are from week class for amino acid 'Ser'. While checking the position of the mutations, we observe that they are from the active site domain of the protein, which may be altered due to mutations and ultimately lead to a



(a) Overall mutated codon (%) for each gene.

(b) Active site mutated codon (%) for each gene.

Figure 4: Gene wise percentage of mutated codons present in each class (Strong, Transitional and Weak) affected due to mutations.

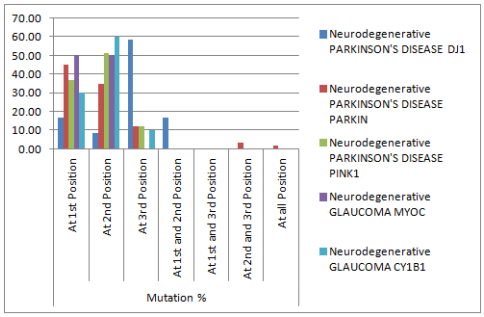


Figure 5: Mutations occurred (%) at different positions of codons.

pathogenic phenotype.

3.3. Positional quantification of mutation in codons

Single point mutation may occur in any random position of codons. In a codon, out of three nucleotides, the first two nucleotides (from $5'$ end) usually determines the type of codons [28]. The role of third nucleotides appears to be very negligible. We observe this fact in the mutated dataset of seven genes too (Figure 5). We observe a tendency of mutations usually happens in 1st and 2nd positions in a codon. Comparatively, mutations in third nucleotide are very less with an exception in case of DJ1. In PARKIN, maximum mutations occur in $1st$ position of a codon. In PINK1, CYP1B1 and MYOC, the mutations majorly occur in the second position of the codon. The observation itself depicts the fact that the importance of the first two positions of a codon is more than that of the third position during mutation.

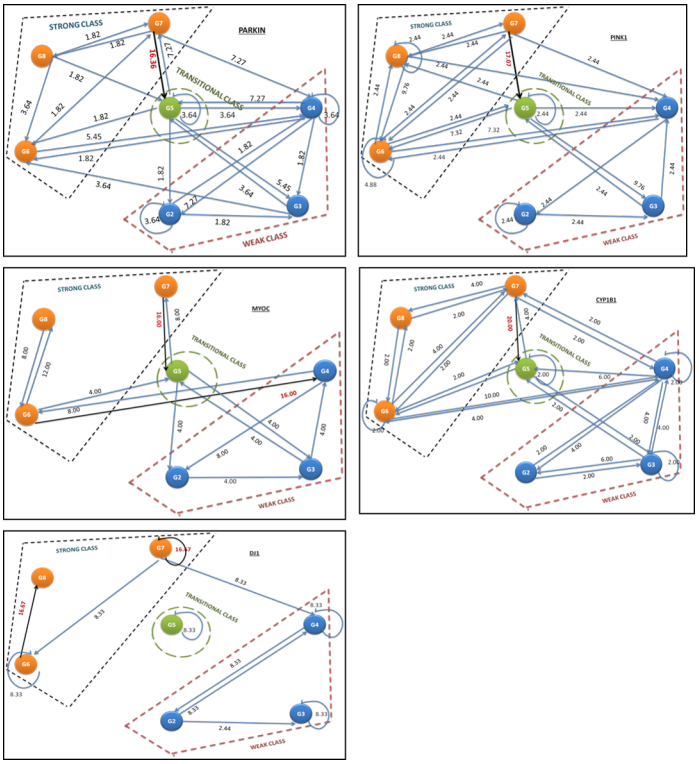


Figure 6: Transition of codon groups during mutations. The highest transitions are highlighted using red colour.

3.4. Transition of codon groups due to mutation

We try to observe any possible change or transition of codon groups among our proposed seven different groups (G2 to G8) during missense mutation. There may be a possibility that during mutations the affected codon may transit from one group to other or may remain unchanged. We compare the changes with normal sequence. The group transition occurs due to mutations for all seven genes that are furnished with a directed graph as shown in Figure 6.

It can be observed that due to the missense mutations in genes responsible for neurodegenerative diseases (both Parkinson's and Glaucoma), majority of mutations occur for G7 group (except DJ1, where we observe both in G7 and G6 group). We further observe that most of the cases G7 group codon transit to G5 group, except DJ1 where G7 group remains unaltered (Supplementary Table S3).It implies that group G7 plays a significant role during mutations

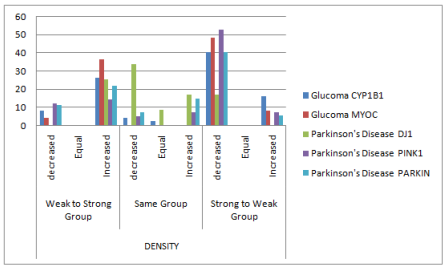


Figure 7: Trend of alteration in density of codon with the change in codon group due to mutation (%)

in neurodegenerative diseases The genes which are responsible for a specific type of disease carry similar types of transition characteristics.

3.5. Alteration of physical density of codon with the change in codon group due to mutation

Due to mutation, the physicochemical properties (density) of the whole gene sequence may be altered. To study the same, quantitatively we consider the major three types of codon classes (Strong, Transitional and Weak) transitions from wild type to disease or mutated. The density of each codon is deducted from the density of nucleotides, where the density of $C = 1.55\ g/cm^3$, $G = 2.2\ g/cm^3$, $T = 1.32\ g/cm^3$ and $A = 1.60\ g/cm^3$. Hence, the density of a codon is the average densities of the constituent nucleotides of the codon. We calculate the change (%) of density as shown in Figure 7. According to our scheme the seven groups of codons maintain the chronological order as $G2 < G3 < G4 < G5 < G6 < G7 < G8$, where $G8$ is the strongest and $G2$ is the weakest group respectively. It is to be noted that the changes in physical density property of codon due to mutations in most of the cases increases (or decreases) with the transitions from weaker group to stronger or stronger group to weaker respectively. The overall analysis depicts that the physical density of codons is proportional to the strength of codons (Supplementary Table S4).

3.6. Investigating proximity between amino acid sequences based on their determinative degrees

In this subsection it is tried to investigate proximity between five genes (namely, PARKIN, PINK1, DJ1, CYP1B1, MYOC) responsible for neurodegenerative diseases based on their determinative degree. To do so a distance matrix has been constructed to calculate the deviations between the

determinative degrees of their wild type amino acid sequences. Let determinative degree of any two arbitrary amino acid sequences S_1 and S_2 are $\delta'(S_1)$ and $\delta'(S_2)$, then the proximity between them with respect to their determinative degree is shown below:

$$\xi(S_1, S_2) = \sum_{n=1}^{20} |\delta'(S_1) - \delta'(S_2)| \tag{4}$$

Table 4: Proximity between amino acid sequences

	PARKIN	PINK1	DJ1	CYP1B1	MYOC
PARKIN	0.00	0.75	23.13	7.05	18.13
PINK1	0.75	0.00	22.39	6.30	17.38
DJ1	23.13	22.39	0.00	16.08	5.01
CYP1B1	7.05	6.30	16.08	0.00	11.08
MYOC	18.13	17.38	5.01	11.08	0.00

The distance matrix shown in table 4 has been constructed using the equation (4) and shows the proximity between the candidate amino acids. A phylogenetic tree is constructed from the distance matrix displayed in table. Thus, although PARKIN, PINK1 and DJ1 genes are responsible for Parkinson's disease and CYP1B1 and MYOC are for Glaucoma, but the pictorial representation of the matrix is able to depict evolutionary relationships between all five genes responsible for neuro-degenerative disease as a whole.

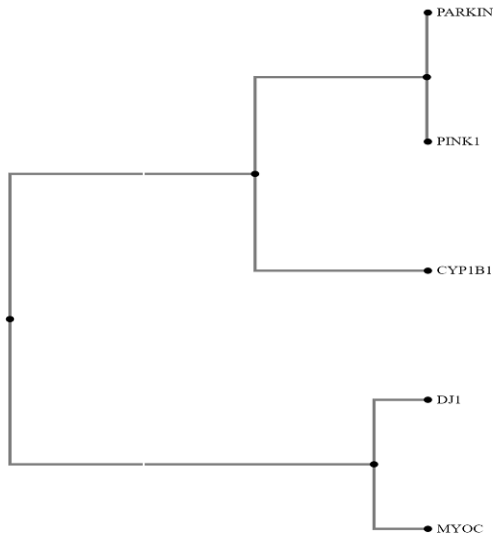
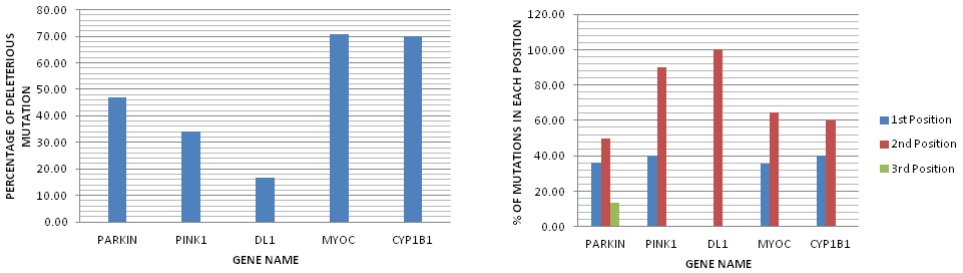


Figure 8: Phylogenetic tree showing evolutionary relationships among all five genes.

3.7. Predicting the functional effect of mutations occurred

Mutations may be amino acid substitution or indel, intensity of harmfulness may not be same always. The mutation may or may not have impact on the biological functions of a protein. PROVEN (Protein Variation Effect Analyzer) is an online software which predicts the functional effect of amino acid substitution. In this part of the manuscript using PROVEN software [29] firstly we find out the percentage of deleterious mutations occurred in each gene.Then it is tried to investigate the percentage of deleterious mutations occurred due to substitutions of nucleotides in different positions of codons. Gene-wise percentage of the deleterious mutations occurred are shown in fig 9(a). The graph shows maximum deleterious mutations taken places in MYOC and the chronological order is MYOC > CYPIB1 > PARK1 > PINK1 > DJ1, where MYOC and CYPIBI both genes are responsible for Glaucoma. Fig 9 (b) shows us the percentage of deleterious mutations occurred due to substitution of nucleotides at three different places of codon. It is remarkable that in all the five genes deleterious mutations occur due to nucleotide substitution at 2nd position of codon. Investigations has also been carried out to find the codon group transitions for deleterious muta-

tions. It is observed that the the deleterious mutations are majorly occurred in strong group of codons. In mutations occurred in the genes of PINK1, DJ1 and CYP1B1 the codon class remains unchanged and in PARKIN and MYOC due to mutations the codons have been changed from strong class to intermediate group (Table 5).



(a) (b)

Figure 9: (a) percentage of deleterious mutations occurred in all the genes. (b) Trend of alteration in different positions of codon(%) causing deleterious mutations

Table 5: Transition of codon groups due to deleterious mutation in percentage

	S2S	S2I	S2W	I2S	I2T	I2W	W2S	W2I	W2W
PARKIN	13.64	36.36	18.18	9.09	0.00	9.09	0.00	0.00	13.64
PINK1	30.77	15.38	7.69	15.38	0.00	0.00	7.69	7.69	15.38
DJ1	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MYOC	17.65	5.88	5.88	11.76	0.00	11.76	11.76	5.88	17.65
CYP1B1	20.00	25.71	14.29	5.71	0.00	0.00	5.71	5.71	22.86

4. Conclusion

In this present work we proposed a classification scheme for codon as well as amino acids based on their determinative degree or strength. Using the classification rule we studied the various interesting codon alteration patterns during missense mutations for two different type of neuro-degenerative diseases. We observed that due to mutation the codons traverse either from strong to transitional and transitional to weak classes or weak to transitional and transitional to strong classes. The overall analysis depicts that the genes which are responsible for a specific type of neuro-degenerative diseases like Parkinson's disease, Glaucoma carry similar types of properties and similar

types of characteristics. The determinative degrees of amino acids depict that the strength or determnative degree of amino acid sequence are always not only depended upon the determinative degree of the constituent amino acids but sometimes (for Arg, Ser, Leu) also the class of the codon which codes for the amino acids. Moreover, calculation of determinative degree of 20 amino acids and getting determinative degree of an amino acid sequence from them may help biologists to understand the evolutionary origin occurrence frequencies of amino acids in proteins. At last it can be concluded that our model gives us more microscopic view of the existing genetic code table and able to provide information in depth when mutations take place. It may also help an individual who is asymptomatic but having the particular genetic change (mutation) for early detection and diagnosis during any terminal diseases. We are planning to study the patterns in other terminal diseases in future.

[1] S. Benzer, et al., The elementary units of heredity., The elementary units of heredity. .

[2] M. Kowalczuk, P. Mackiewicz, D. Szczepanik, A. Nowicka, M. Dudkiewicz, M. Dudek, S. Cebrat, Multiple base substitution corrections in DNA sequence evolution, International Journal of Modern Physics C 12 (07) (2001) 1043–1053.

[3] J. K. Das, R. Singh, P. P. Choudhury, B. Roy, Identifying Driver Potential in Passenger Genes Using Chemical Properties of Mutated and Surrounding Amino Acids, in: Computational Intelligence and Big Data Analytics, Springer, 107–118, 2019.

[4] D. P. Minde, Z. Anvarian, S. G. Rüdiger, M. M. Maurice, Messing up disorder: how do missense mutations in the tumor suppressor protein APC lead to cancer?, Molecular cancer 10 (1) (2011) 101.

[5] M. Dermit, M. Dodel, F. K. Mardakheh, Methods for monitoring and measurement of protein translation in time and space, Molecular BioSystems 13 (12) (2017) 2477–2488.

[6] C. R. Woese, D. Dugre, S. Dugre, M. Kondo, W. Saxinger, On the fundamental nature and evolution of the genetic code, in: Cold Spring Harbor symposia on quantitative biology, vol. 31, Cold Spring Harbor Laboratory Press, 723–736, 1966.

[7] U. Lagerkvist, " Two out of three": an alternative method for codon reading, Proceedings of the National Academy of Sciences 75 (4) (1978) 1759–1762.

[8] T. Négadi, The Genetic Codes: Mathematical Formulae and an Inverse Symmetry-Information Relationship, Information 8 (1) (2016) 6.

[9] Y. B. Rumer, Translation of 'Systematization of codons in the genetic code [I]'by Yu. B. Rumer (1966), Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2063) (2016) 20150446.

[10] A. Sengupta, J. K. Das, P. P. Choudhury, Investigating evolutionary relationships between species through the light of graph theory based on the multiplet structure of the genetic code, in: 2017 IEEE 7th International Advance Computing Conference (IACC), IEEE, 854–859, 2017.

[11] J. K. Das, A. Sengupta, P. P. Choudhury, S. Roy, Mapping sequence to feature vector using numerical representation of codons targeted to amino acids for alignment-free sequence analysis, Gene 766 (2021) 145096.

[12] A. Sengupta, P. P. Choudhury, S. Chakraborty, Sequence Characterization of Glutamate Receptor Genes of Rat (Vertebrate) and Arabidopsis Thaliana (Plant), in: Proceedings of International Conference on Frontiers in Computing and Systems, Springer, 495–510, 2018.

[13] D. L. Gonzalez, The mathematical structure of the genetic code, in: The Codes of Life, Springer, 111–152, 2008.

[14] D. L. Gonzalez, S. Giannerini, R. Rosa, et al., The mathematical structure of the genetic code: a tool for inquiring on the origin of life, Statistica 69 (3-4) (2009) 143–157.

[15] D. Gonzalez, S. Giannerini, R. Rosa, Circular codes revisited: a statistical approach, Journal of Theoretical Biology 275 (1) (2011) 21–28.

[16] T. H. Jukes, T. H. Jukes, Molecules and evolution, Tech. Rep., Columbia University Press New York, 1966.

[17] D. L. Gonzalez, S. Giannerini, R. Rosa, Rumer's transformation: A symmetry puzzle standing for half a century, Biosystems 187 (2020) 104036.

[18] V. Karasev, V. Stefanov, Topological nature of the genetic code, Journal of theoretical biology 209 (3) (2001) 303–317.

[19] D. Duplij, S. Duplij, Determinative degree and nucleotide content of DNA strands, Ile 51 (3) (2001) 3.

[20] J. K. Das, P. P. Choudhury, A. Chaudhuri, S. S. Hassan, P. Basu, Analysis of Purines and Pyrimidines distribution over miRNAs of Human, Gorilla, Chimpanzee, Mouse and Rat, Scientific reports 8 (1) (2018) 9974.

[21] J. M. Berg, Biochemistry 5th Edition, 2006.

[22] D. Gilis, S. Massar, N. J. Cerf, M. Rooman, Optimality of the genetic code with respect to protein stability and amino-acid frequencies, Genome biology 2 (11) (2001) 1–12.

[23] S. Itzkovitz, U. Alon, The genetic code is nearly optimal for allowing additional information within protein-coding sequences, Genome research 17 (4) (2007) 405–412.

[24] S. Hormoz, Amino acid composition of proteins reduces deleterious impact of mutations, Scientific reports 3 (1) (2013) 1–10.

[25] A. Abeliovich, A. D. Gitler, Defects in trafficking bridge Parkinson's disease pathology and genetics, Nature 539 (7628) (2016) 207–216.

[26] S. J. Annesley, S. T. Lay, S. W. De Piazza, O. Sanislav, E. Hammersley, C. Y. Allan, L. M. Francione, M. Q. Bui, Z.-P. Chen, K. R. Ngoei, et al., Immortalized Parkinson's disease lymphocytes have enhanced mitochondrial respiratory activity, Disease models & mechanisms 9 (11) (2016) 1295–1305.

[27] H.-W. Wang, P. Sun, Y. Chen, L.-P. Jiang, H.-P. Wu, W. Zhang, F. Gao, Research progress on human genes involved in the pathogenesis of glaucoma, Molecular medicine reports 18 (1) (2018) 656–674.

[28] F. Lustig, P. Elias, T. Axberg, T. Samuelsson, I. Tittawella, U. Lagerkvist, Codon reading and translational error. Reading of the glutamine and lysine codons during protein synthesis in vitro., Journal of Biological Chemistry 256 (6) (1981) 2635–2643.

[29] Y. Choi, A. P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (16) (2015) 2745–2747.

Competing interests

The authors declare no competing interests.

Supporting information

S1 Table.. **S1.** The active domains that altered due to mutations.

paragraph*S1 Table. **S2.** Mapping between the seven codon groups to chemical groups according to the chemical properties of an amino acids.

paragraph*S1 Table. **S3.** Transition of codon groups during mutation. The highest transitions are highlighted using red colour.

paragraph*S1 Table. **S4.** Trend of alteration in density of codon with the change in codon group due to mutation (%).