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

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

[a]*Department of Master of Computer Applications, MCKV Institute of Engineering, Liluah, 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*

1. Introduction

Genes are the functional units of heredity [4]. 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. Mutation differs from Single Nucleotide Polymorphism(SNP) in many ways [23]. For instance, occurrence of mutation in a population should be less than 1% whereas SNP occurs with greater than 1%. Mutation always occurs in diseased group whereas SNP is occurs in both diseased and control population. Mutation is responsible for some disease phenotype but SNP may or may not be associated with disease phenotype. Mutations occur in two ways: i) a base

*Current address: Department of Pediatrics, Johns Hopkins University School of Medicine, Maryland, USA
**Corresponding author

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. Researchers predicting that SNPs influence the susceptibility of complex diseases [17, 24, 5, 22, 12, 2]. Base substitutions can have a variety of effects [14]. 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 [9, 18]. 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 [10].

The alteration of amino acid due to the alternation in codon during disease may not be an arbitrary phenomenon [27]. 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 [15, 19, 20, 21]. 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. 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 calculate determinative degree of each codon followed by degree of each amino acid. 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[13] and their effects in the physical and chemical properties at DNA primary sequential level as well as secondary structural level.

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 [15]. Rumer [20] 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.

In order to represent a codon uniquely, we introduce a positional specification of each doublet or root in the rhombic organization of roots. 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, Duplij et. al. [11] 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 as shown in Figure 1.

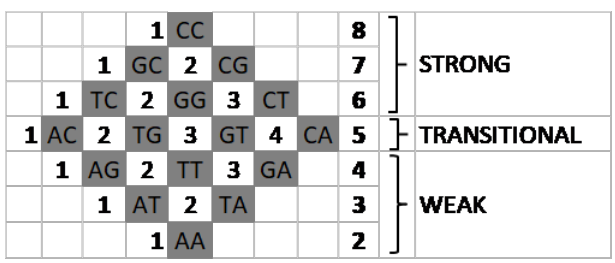


Figure 1: Rhombic structure of roots. Order of three dual nucleotide properties (Bottom to top): Week, Transitional and Strong. The positional specifications of each dual nucleotide are recorded from left to right).

Using the numerical scores of each dual nucleotides in above classification and their positions in the Rumer's rhombic arrangement, we represent each codon using three digits number as shown in Figure 2. 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 second digit represents the position of the root in a particular row in the rhombic structure. The last digit is the determinative degree of the third nucleotide. 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. We termed it 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=Group or Row number in rhombic structure, D_2=Column number in rhombic structure and D_3 is the determinative degree of third nucleotides as discussed above.

Based on this scheme, we may code the entire codon table as given in Table 1.

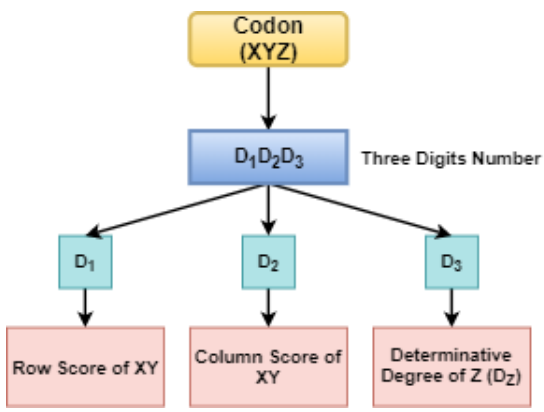


Figure 2: Scheme for three digits numerical representation of codons based on dual nucleotide scores and degenerative degree of third nucleotide.

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 **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 [7]. We classify all 64 codons into three classes namely Weak, Strong and Transitional according to

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

	T=2		C=4		A=1		G=3		
	Codon	AA	Codon	AA	Codon	AA	Codon	AA	
T=2	422	F	612	S	322	Y	522	C	T=2
	424	F	614	S	324	Y	524	C	C =4
	421	L	611	S	321	STOP	521	STOP	A =1
	423	L	613	S	323	STOP	523	W	G =3
C=4	632	L	812	P	542	H	722	R	T =2
	634	L	814	P	544	H	724	R	C =4
	631	L	811	P	541	Q	721	R	A =1
	633	L	813	P	543	Q	723	R	G =3
A=1	312	I	512	T	212	N	412	S	T =2
	314	I	514	T	214	N	414	S	C =4
	311	I	511	T	211	K	411	R	A =1
	313	M	513	T	213	K	413	R	G =3
G=3	532	V	712	A	432	D	622	G	T =2
	534	V	714	A	434	D	624	G	C =4
	531	V	711	A	431	E	631	G	A =1
	533	V	713	A	433	E	643	G	G =3

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. The overlapping distribution of all the amino acids across three different classes are shown with the help of Ven-diagram in Figure 3.

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

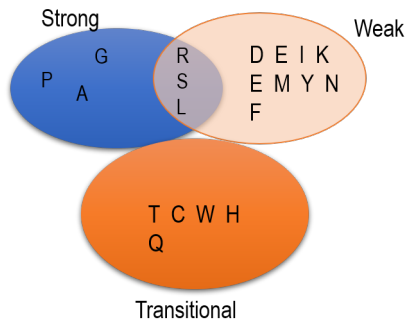


Figure 3: Distribution of 20 amino acids in three different overlapping classes.

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.

2.4. Codon classes and their chemical groupings

The structure of a protein is depended upon the chemical property of the amino acid by which it is formed. We classified 20 amino acids into seven (07) groups (G2 to G8) based on the determinative degree of the constituent

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	Weak Root	212	Ser	Weak Root	413
Asn		213	Leu		422
Ile		312.33	Phe		423
Met		313	Glu		432
Tyr		323	Asp		433
Arg		412			
Thr	Transitional Root	512.5	Ser	Strong Root	612.5
Cys		523	Gly		630
Trp		523	Leu		632.5
Val		532.5	Ala		712.5
Gln		542	Arg		722.5
His		543	Pro		812.5

condon. We try to make a mapping between the seven codon groups to chemical groups according to the chemical properties of an amino acids and reported in Supplementary Table TS2 . We focus mainly on two different chemical properties of amino acids. The side chain structure (R-group) based eight (08) chemical groups [8, 6] and polarity based two groups [25]. Polarity is a separation of electric charge leading to a molecule or its chemical groups having an electric dipole moment, with an end having negatively charge and another end with positively charged.

2.5. A new codon wheel representation

Based on our proposed scheme we represent each amino acids and their chemical classes in a wheel form for the ease of finding the numerical value of any amino acid as per our scheme. The codon wheel is shown in Figure 4. To use the codon wheel, start from the center with 1(A), 2(T), 3(G), 4(C), i.e. layer 1. Thus follow the wheel from center to outer layers. In second layer of the wheel, 16 possible dual nucleotides are kept. Those are the nucleotides occupying the first and second positions of codon and according to our methodology their additive degree is occupying the first position of our codon format mentioned in Figure2. Layer 3 is basically stating the positions of those dual nucleotides in rhombic structure at Figure 4 and occupying second position of codon. At layer 4 the four bases are assigned for each dual nucleotides, making the places of nucleotides at third positions of codons. Thus at layer 5 we get the 64 codons. Next, translate the codons into an amino acid from the mRNA codons. This process is called as RNA translation. Once completed, follow the RNA sequence to find the amino acids that it translates to get protein finally. At layer 6 the amino acids

are stated which can be coded by codons specified at layer 5 respectively. At layer 7 based on the determinative degrees of codons, the corresponding groups are assigned and thus the amino acids form groups too according to the groups of codons by which the have been coded. Finally the 64 codons along with the 20 standard amino acids and 3 stop codons are classified into 3 different classes (Strong(S), Weak (W), and Transitional (T)) at layer 8. As a example, for codon CCA, starting from center, nucleotide 4(C) at first layer will go to a dual nucleotide CC (4+4=8) at layer 2. Thus getting the numeric value 8 at first position of codon, it will go to layer 3 to get the numeric value at second position of codon, which is 1 as per rhombic structure specified in Figure 1. So now at the end of layer 3 the value is 81. At layer 4 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, given at layer 5. It codes for the amino acid P(Proline), which is specified at layer 6. As according to the classification of dual nucleotide the CC comes under group G8 and class of Strong(S), so layer 7 is indicating the group and layer 8 is indicating the class respectively. At last, the layer 9 reflects the chemical properties which the amino acids belong to.

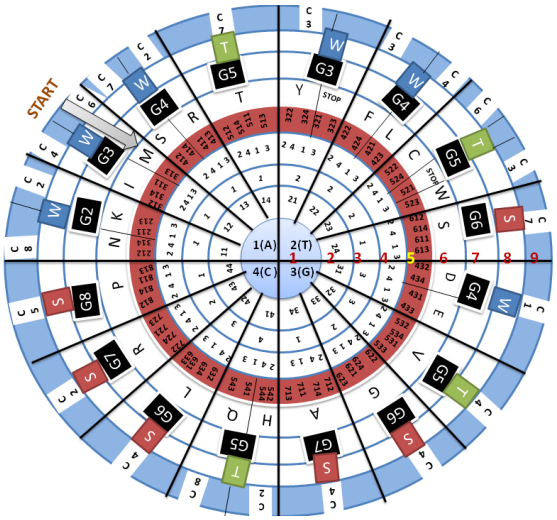


Figure 4: The codon wheel structure: organization of dual nucleotides, codons and amino acid chemical properties.

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 four types of genetic diseases namely, Parkinson, Glaucoma, Oculocutaneous Albinism type 1 and Wilson's Disease for experimentation. Among them, Parkinson's Disease and Gloucoma are neurodegenerative diseases, and Oculocutaneous Albinism type 1 and Wilson's Diseases are monogenic diseases. Parkinson's disease is a progressive disease of the nervous system, where PINK1, PARKIN and DJ1 are three main pathogenic genes [1, 3]. Glaucoma is a condition of increased ocular pressure within the eye, causing gradual loss of sight [26]. MYOC and CYP1B1 are two significant genes responsible for occurring this disease. Oculocutaneous albinism type 1 is a disease that affects the coloring of the skin, hair, and eyes. Mutations in TYR (Tyrosinase gene) leads for this disease. Mutation in ATP7B gene leads to Wilson disease prevents the body from removing extra copper, causing copper to build up in the liver, brain, eyes, and other organs. We took a group of mutations reported in these genetic diseases collected from various online sources[1],[2],[3], [4], [5].

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

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

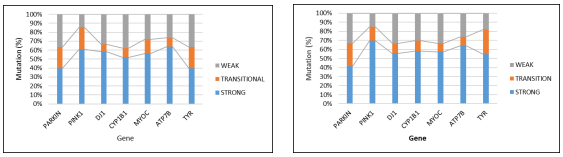
Table 3: Mutation datasets (Missense sample (MS) and nonsense sample (NS)) of four disease types and seven genes used for experimentation.

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	
Monogenic	Oculocutaneous Albinism type 1	TYR	50	
	Wilson's Disease	ATP7B	50	8

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 5. 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 TS1. 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 5(b)). 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 and TYR, 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 pathogenic phenotype.



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

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

Figure 5: Gene wise percentage of mutated codons present in each class (Strong, Transitional and Weak) affected due to mutations.

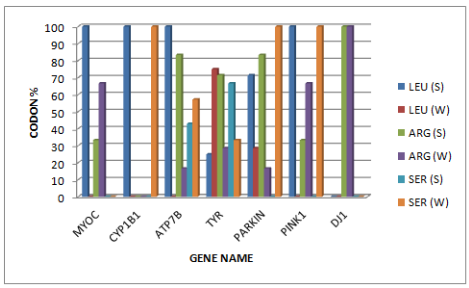


Figure 6: Mutated codon (%) in Week (W) and Strong (S) codon classes for the three amino acids: Ser, Arg and Leu.

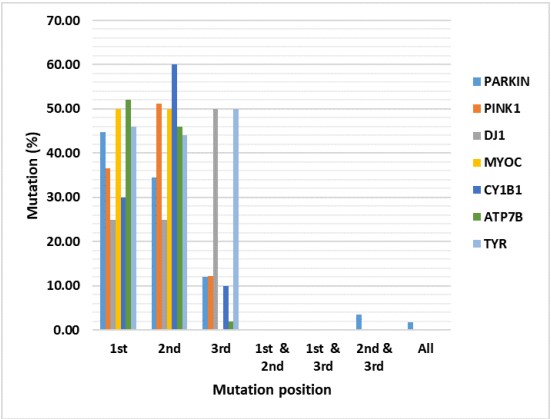


Figure 7: Mutations occurred (%) at different positions of codons.

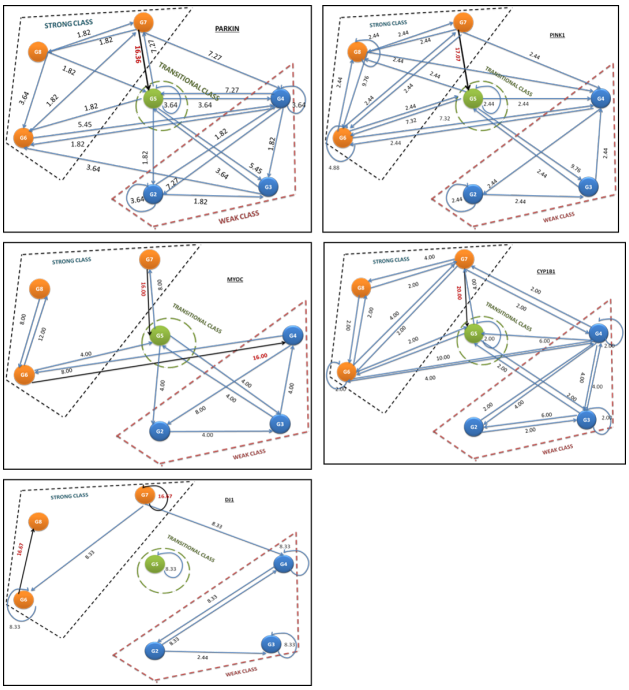
3.3. Positional quantification of mutation in codons

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 [16]. The role of third nucleotides appears to be very negligible. We observe this fact in the mutated dataset of seven genes too (Figure 7). 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.

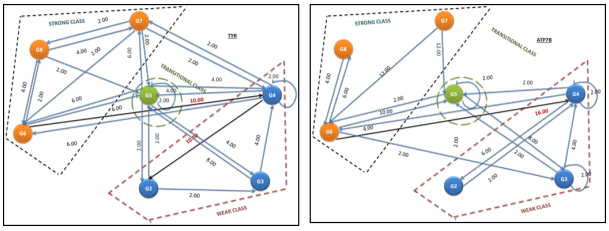
3.4. Transition of codon groups due to mutation

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 8.

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

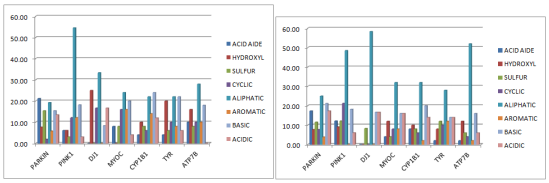


(a) Neurodegenerative Disease



(b) Monogenic Disease

Figure 8: Transition of codon groups during mutation. The highest transitions are highlighted using red colour.



(a) Before mutations. (b) After mutations.

Figure 9: Changes in chemical properties of amino acids: (a) Before mutations and (b) After mutations.

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 TS3). But in monogenic diseases i.e. Wilson's Disease and Oculocutaneous albinism type 1, major changes occurs in group G6 and transit group G4 during mutation. It implies that group G7 plays a significant role during mutations in neurodegenerative diseases and G6 in monogenic diseases. The genes which are responsible for a specific type of disease carry similar types of transition characteristics.

3.5. Transition of chemical properties of amino acids due to mutation

Missense mutation refers to the change of amino acid. Therefore, it may change the chemical groups too. In this section, we try to study the transition of chemical groups for amino acids during mutation.

We compare the distribution of chemical groups for normal and candidate missense mutated sequences and reported in Figure 9. We observe an interesting fact that in missense mutations, significant changes occur in aliphatic group. However, no such changes have been observed in nonsense mutation.

We also study the transition of acidic property (polar and non polar) during mutation and report the transition graph in Figure 10. In case of genes responsible for monogenic disease i.e. ATP7B and TYR, most of the amino acids affected during mutation are polar and non-polar respectively. Interestingly after the mutation, the majority of them changed to another amino acid, but within the same (i.e. non-polar) acidity group. We do not observe any significant trends for the genes responsible for neurodegenerative diseases.

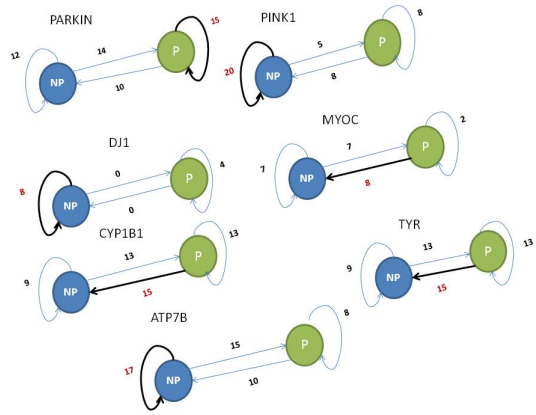


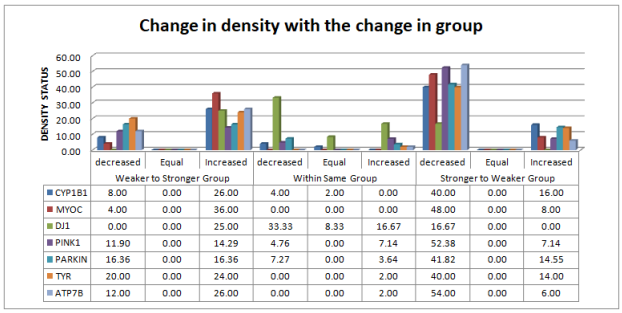
Figure 10: Changes in Chemical Properties (Polar vs. Non-polar) of Amino Acids.

3.6. 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 11. 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 TS4).

4. Conclusion

We proposed a classification scheme for codon as well as amino acids based on their determinative degree. Using the classification we studied the



	Weaker to Stronger Group			Within Same Group			Stronger to Weaker Group		
	decreased	Equal	Increased	decreased	Equal	Increased	decreased	Equal	Increased
CYP1B1	8.00	0.00	26.00	4.00	2.00	0.00	40.00	0.00	16.00
MYOC	4.00	0.00	36.00	0.00	0.00	0.00	48.00	0.00	8.00
DJ1	0.00	0.00	25.00	33.33	8.33	16.67	16.67	0.00	0.00
PINK1	11.90	0.00	14.29	4.76	0.00	7.14	52.38	0.00	7.14
PARKIN	16.36	0.00	16.36	7.27	0.00	3.64	41.82	0.00	14.55
TYR	20.00	0.00	24.00	0.00	0.00	2.00	40.00	0.00	14.00
ATP7B	12.00	0.00	26.00	0.00	0.00	2.00	54.00	0.00	6.00

Figure 11: Trend of alteration in density of codon with the change in codon group due to mutation (%)

various interesting codon alteration patterns during missense mutations for two different type of 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 disease carry similar types of properties and similar types of characteristics. 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.

References

[1] Asa Abeliovich and Aaron D Gitler. Defects in trafficking bridge parkinson's disease pathology and genetics. Nature, 539(7628):207–216, 2016.

[2] Hiroshi Akashi. Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. Genetics, 136(3):927–935, 1994.

[3] Sarah J Annesley, Sui T Lay, Shawn W De Piazza, Oana Sanislav, Eleanor Hammersley, Claire Y Allan, Lisa M Francione, Minh Q Bui, Zhi-Ping Chen, Kevin RW Ngoei, et al. Immortalized parkinson's disease

lymphocytes have enhanced mitochondrial respiratory activity. Disease models & mechanisms, 9(11):1295–1305, 2016.

[4] Seymour Benzer et al. The elementary units of heredity. The elementary units of heredity., 1956.

[5] Anthony J Brookes. The essence of snps. Gene, 234(2):177–186, 1999.

[6] Jayanta Kumar Das and Pabitra Pal Choudhury. Chemical property based sequence characterization of ppca and its homolog proteins ppcb-e: A mathematical approach. PloS one, 12(3):e0175031, 2017.

[7] Jayanta Kumar Das, Pabitra Pal Choudhury, Adwitiya Chaudhuri, Sk Sarif Hassan, and Pallab Basu. Analysis of purines and pyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat. Scientific reports, 8(1):9974, 2018.

[8] Jayanta Kumar Das, Provas Das, Korak Kumar Ray, Pabitra Pal Choudhury, and Siddhartha Sankar Jana. Mathematical characterization of protein sequences using patterns as chemical group combinations of amino acids. PloS one, 11(12):e0167651, 2016.

[9] Jayanta Kumar Das, Richa Singh, Pabitra Pal Choudhury, and Bidyut Roy. Identifying driver potential in passenger genes using chemical properties of mutated and surrounding amino acids. In Computational Intelligence and Big Data Analytics, pages 107–118. Springer, 2019.

[10] Maria Dermit, Martin Dodel, and Faraz K Mardakheh. Methods for monitoring and measurement of protein translation in time and space. Molecular BioSystems, 13(12):2477–2488, 2017.

[11] Diana Duplij and Steven Duplij. Determinative degree and nucleotide content of dna strands. Ile, 51(3):3, 2001.

[12] Hila Gingold and Yitzhak Pilpel. Determinants of translation efficiency and accuracy. Molecular systems biology, 7(1), 2011.

[13] VA Karasev and VE Stefanov. Topological nature of the genetic code. Journal of theoretical biology, 209(3):303–317, 2001.

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

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

[16] F Lustig, P Elias, T Axberg, T Samuelsson, I Tittawella, and 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):2635–2643, 1981.

[17] Eden R Martin, Eric H Lai, John R Gilbert, Allison R Rogala, AJ Afshari, John Riley, KL Finch, JF Stevens, KJ Livak, Brandon D Slotterbeck, et al. Snping away at complex diseases: analysis of single-nucleotide polymorphisms around apoe in alzheimer disease. The American Journal of Human Genetics, 67(2):383–394, 2000.

[18] David P Minde, Zeinab Anvarian, Stefan GD Rüdiger, and Madelon M Maurice. Messing up disorder: how do missense mutations in the tumor suppressor protein apc lead to cancer? Molecular cancer, 10(1):101, 2011.

[19] Tidjani Négadi. The genetic codes: Mathematical formulae and an inverse symmetry-information relationship. Information, 8(1):6, 2016.

[20] Yu 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):20150446, 2016.

[21] Antara Sengupta, Jayanta Kumar Das, and Pabitra Pal 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), pages 854–859. IEEE, 2017.

[22] C Smith. Genomics: Snps and human disease. Nature, 435(7044):993, 2005.

[23] Lingtao Su, Guixia Liu, Han Wang, Yuan Tian, Zhihui Zhou, Liang Han, and Lun Yan. Research on single nucleotide polymorphisms interaction detection from network perspective. PLoS One, 10(3):e0119146, 2015.

[24] Yousin Suh and Jan Vijg. Snp discovery in associating genetic variation with human disease phenotypes. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 573(1-2):41–53, 2005.

[25] MV Volkenstein. Coding of polar and non-polar amino-acids. Nature, 207(4994):294–295, 1965.

[26] Hong-Wei Wang, Peng Sun, Yao Chen, Li-Ping Jiang, Hui-Ping Wu, Wen Zhang, and Feng Gao. Research progress on human genes involved in the pathogenesis of glaucoma. Molecular medicine reports, 18(1):656–674, 2018.

[27] Carl R Woese, DH Dugre, SA Dugre, M Kondo, and WC Saxinger. On the fundamental nature and evolution of the genetic code. In Cold Spring Harbor symposia on quantitative biology, volume 31, pages 723–736. Cold Spring Harbor Laboratory Press, 1966.