Probability of Change in Life: amino acid changes in single nucleotide substitutions

Kwok-Fong Chan[1,†], Stelios Koukouravas[1,†], Joshua Yi Yeo[1], Darius Wen-Shuo Koh[1], Samuel Ken-En Gan[1,*]

[1] Antibody & Product Development Lab, BII, A*STAR, Singapore 138671

† Both authors contributed equally to this work

* Corresponding Author: Email: samuelg@bii.a-star.edu.sg

Bioinformatics Institute, A*STAR

30 Biopolis Street, #07-01 Matrix

Singapore 138671

Tel: +65 6407 0584

ABSTRACT

Mutations underpin the processes in life, be it beneficial or detrimental. While mutations are assumed to be random in the bereft of selection pressures, the genetic code has underlying computable probabilities in amino acid phenotypic changes. With a wide range of implications including drug resistance, understanding amino acid changes is important. In this study, we calculated the probabilities of substitutions mutations in the genetic code leading to the 20 amino acids and stop codons. Our calculations reveal an enigmatic in-built self-preserving organization of the genetic code that averts disruptive changes at the physicochemical properties level. These changes include changes to start, aromatic, negative charged amino acids and stop codons. Our findings thus reveal a statistical mechanism governing the relationship between amino acids and the universal genetic code.

Keywords: Codon single base mutation; Single nucleotide substitution; Probability; Amino acid; Mutation

INTRODUCTION

Protein translation decodes DNA in the frame of three bases at a time referred to as a codon (Crick *et al.*, 1961). The open reading frame (ORF) typically begins with a Kozak sequence (Kozak, 1984) embedding the start codon (ATG, coding for Methionine) and ends with a stop codon (TAA, TAG or TGA). The codons are translated into amino acids based on the universal genetic code (Crick, 1968) with rare exceptions. The majority of the common 20 amino acids are encoded by multiple codons (Crick *et al.*, 1961; Lagerkvist, 1978) in the genetic code, which is degenerative at the third base, described as the Wobble Hypothesis (Crick, 1966).

Mutations in DNA underpin many life processes and can occur as insertion, deletion and single nucleotide substitutions (SNS). Occurring at varying rates in life processes that range from hypermutations in the immune system e.g. antibodies (Roth & Craig, 1998; Su *et al.*, 2017; Ling *et al.*, 2018) to disease development in cancer (Hollstein *et al.*, 1991; Hanahan & Weinberg, 2011), and drug resistance (Weiss, 1993; Su *et al.*, 2016; Chiang *et al.*, 2018), mutations underpin change. Given that insertions and deletions cause frameshifts that are often detrimental in most cases, SNS are generally more frequent and are categorized into missense, nonsense and silent. Such SNS types often lead to disease states that include the non-conservative missense mutation in the β-globin gene from Glutamate (GAG) into Valine (GTG) in sickle cell anaemia (The International F.M.F. Consortium, 1997). On the other hand, nonsense mutations to stop codons underlie cystic fibrosis (Tsui, 1992) and beta-thalassemia (Cao & Galanello, 2010), where the introduction of a premature stop codon truncates key proteins to result in a loss of function.

Analysing the genetic code table, we found clear biases to the changes that mutations can achieve in limited mutation events. It is virtually impossible for

Methionine (ATG) to mutate to a codon encoding for Proline (CCA, CCC, CCA, CCG) in a single SNS mutational event even though it is possible to become a Lysine (AAG) or Leucine (TTG). Such observations hint of an in-built probabilistic predisposition of specified codons to mutate to certain amino acids, demonstrating innate mutational constraints. To address this, we analysed the probability of SNS induced changes in all 64 codons, with the aim of calculating the probable change outcomes of each codon.

MATERIALS & METHODS

Base probability of amino acid in the genetic code

In calculating the base probability of single substitution mutations (T, A, C, or G, respectively) occurring at the specific codon location to lead to a specific amino acid was calculated as follows in Equation 1:

$$Base\ Probability\ of\ amino\ acid = \frac{Total\ No.of\ codons\ encoding\ AA}{Total\ permutations\ of\ mutations\ in\ the\ genetic\ code} \qquad \textbf{Equation 1}$$

The total number of possible mutations that can occur in the 64 codons is 144 (i.e. 3 codon positions x 64 codons, then excluding 48 possible positions for each base calculated based on 16 permutations for the specific base at the respective 3 codon positions). For example, the probability of Tyrosine (TAC or TAT) if a SNS of T occurs in the codon at the first position from **C**AT, **C**AC, **A**AT, **A**AC, **G**AT, **G**AC, or at the last position from TA**C**, TA**A**, TA**G** would be $\frac{9}{144} = 0.63$.

If the target base (T, A, G or C) to be mutated to is left undefined, the total combination becomes 64 codons x 3 codon positions x 4 bases = 768 possible permutations. Excluding mutations to self, there are 64 codons x 3 codon positions = 192. This brings the possible mutations to 768 – 192 = 576. For Tyrosine, the number of codons of all bases becomes 18 as follows – T : { **C**AT, **C**AC, **A**AT, **A**AC, **G**AT, **G**AC, TA**C**, TA**A**, TA**G** }, A : { T**T**T, T**T**C, T**C**T, T**C**C, T**G**T, T**G**C }, C : { TA**T**, TA**A**, TA**G** } with a final combined total probability of $\frac{18}{576} = 0.031$

Probability of amino acids/stop codons changes

In the probability of amino acid change, the number of possible changes of the total codon for a specific amino acid is the number of codons multiplied by 9 (reflecting the other 3 bases in the 3 positions of the codon) as shown in Equation 2.

$$Probability\ of\ amino\ acid\ change = \frac{No.\ of\ codons\ of\ specific\ amino\ acid}{No. of\ possible\ changes\ in\ the\ total\ codons\ for\ the\ specific\ amino\ acid} \quad \textbf{Equation 2}$$

For example, for Arginine with 6 codons, there would be 6 x 9 = 54 possible changes in the total codons. Considering the mutations that can lead to Proline from Arginine codons as C**G**T, C**G**C, C**G**A, C**G**G, the probability of Arginine mutating to Proline would be $\frac{4}{54} = 0.074$

Probability of Physicochemical property changes

When grouped into the physicochemical groups, the number of possible changes in the codons of the amino acids belonging to the specific physicochemical property group would be the denominator and calculated for specific groups using equation 3.

$$Probability\ of\ physicochemical\ property\ change = \frac{No.\ of\ codons\ of\ amino\ acids\ in\ physicochemical\ group}{No.\ of\ possible\ changes\ in\ the\ codons\ in\ the\ physicochemical\ group} \quad \boldsymbol{Equation\ 3}$$

For example, in the negatively charged amino acid group, there is Aspartate and Glutamate encoded by GAT, GAC and GAA, GAG, respectively. With the possibility of changes to the other 3 bases in the 3 locations of the codon to form the total possible changes becomes 3 bases x 3 locations x 4 bases = 36 for the denominator.

For the codons specifying amino acids in the specific physicochemical group, T: { GA**C**, GA**A**, GA**G** }, C : { GA**T**, GA**A**, GA**G** }, A : { GA**T**, GA**C**, GA**G** }, G : { GA**G**, GA**C**, GA**A** } forms 12. Thus, the probability of Aspartate and Glutamate mutating to other codons encoding Aspartate and Glutamate would be $\frac{12}{36} = 0.33$.

RESULTS

We found in-built predispositions in the genetic code when analysing the base probability of the 20 amino acids and stop codons occurring, unpinned by the four bases in specific locations of the codon (see Supplementary Figure S1 and S2A–C). In the example of a T in the first position of a codon, there is a bias towards codons encoding Serine (p = .25) over other amino acids (p < .2) in Figure S1.

Studying the probability of a specific amino acid occurrence in the genetic code, predisposed changes to T, A, C or G lead to Leucine (p = .188), Lysine/Threonine (p = .104), Proline (p = .188) and Glycine (p =.188), respectively (Supplementary Figure S1). This is due to the dominance of specific bases in the respective amino acid codons (Supplementary Table S3). Amino acids such as Serine, with a more balanced usage of the four bases, have more even out probabilities compared with Leucine of also 6 codons. Alternatively, amino acid codons with a heavy bias towards specific bases, such as Proline (towards C) would have a higher probability in a base C mutation event.

Statistically, mutations towards T and A have 18 possible amino acid changes (including the stop codon), whereas C and G substitutions have 16 change possibilities (Figure 1). When re-grouped into transitions (A ↔ G) and transversions (C ↔ T) mutations, the various base mutations have equal possibilities.

Calculating the probabilities of the 20 amino acids and stop codon to mutate to one another, there were interesting patterns observed when calculating the specific base changes (Supplementary Figure S4A-D). In the event of single G substitutions, Glycine will only mutate to itself. Such self-preservation probabilities are also observed for Phenylalanine in T substitutions, Lysine in A substitutions and Proline in C substitutions (Supplementary Figure S4A-D).

Probabilities of change to other amino acids were also not uniform (Supplementary Figure S4A–D and S5A–D). T mutations have unique bias towards aromatic amino acids avoiding Glutamine, Lysine and Glutamate (Supplementary Figure S4A, Supplementary Figure S5A). Mutation to A, on the other hand, have higher probabilities to become stop codon and are less likely to lead to non-polar amino acids than mutations of other bases. Mutations to A also predisposes towards polar

neutral or polar positive amino acids such as Lysine and Asparagine while avoiding Phenylalanine, Tryptophan and Cysteine (Supplementary Figure S4B, Supplementary Figure S5B). C mutations avert both start or stop codons, Lysine, Glutamate and Tryptophan with a strong bias for Proline (Supplementary Figure S4C, Supplementary Figure S5C), and G mutations have strongly pronounced biases towards Glycine and Arginine while avoiding Tyrosine, Isoleucine, Phenylalanine, Asparagine and Histidine (Supplementary Figure S4D, Supplementary Figure S5D).

When studying all 4 base SNS together, there is a pattern of self-preservation shown in the higher probabilities diagonally in Figure 2 of no amino acid change. Self-bias was expectedly absent in amino acids encoded by 1 codon such as Methionine and Tryptophan.

Although this relationship is not linear, codon diversity plays a role in the probability of self-preservation; those amino acids with a more diverse range of codons have a decreased likelihood of remaining self e.g. Arginine and Leucine with 6 codons, have the same self-preservation probability as the amino acids with 4 codons. On the other hand, Serine being more diverse despite also having 6 codons, has reduced self-preserving probability compared to Arginine and Leucine. Given that stop codons can tolerate changes in the second and third codon positions (TAA, TAG and TGA), the stop codons had lower probabilities of self-preservation as compared to Isoleucine of also 3 codons with only changes in the third codon position. (ATT, ATC, ATA) (Figure 2).

When the amino acids are grouped into their physicochemical types, there are self-biases towards silent and conservative changes within the same amino acid group (Figure 3). This self-bias applies to non-polar, polar and positive; the other groups have more levelled probabilities for non-conservative changes. Apart from stop

codon self-preservation at p =.148 (Figure 3), the next highest pre-deposition towards stop codon (nonsense mutation) was by the aromatic group at p=.133. Aromatic amino acids (especially Tryptophan and Tyrosine) have the highest probabilities to a nonsense mutation stop codon at p = .222 individually (Figure 2). Despite the large number of codons, non-polar amino acids have extremely low probabilities to change to a stop codon at p = .017 (Figure 3). Thus, intrinsic barriers against changes of amino acid and at the physicochemical levels are both at the nucleic acid level.

DISCUSSION

The genetic code of life has in-built intrinsic biases and barriers. Within single mutational events, there are fixed probabilities for the type of change in selection pressure-free conditions. Generally, C and G mutations result in less amino acids changes (16 possible changes) than A and T mutations (18 possible changes) (Figure 1).

Within the finite mutational events that can be incorporated into a population (Haldane, 1935), the rate of diverse disruptive changes is further constrained by the mutational bias towards no change, within the physicochemical group, or a slight trend towards non-polar amino acid codons (Figure 3). Such predispositions at the nucleotide level, supports the proposal of the rarity of the genetic code (Freeland & Hurst, 1998) in the universe, especially given the organisation of self-bias statistically.

While real-life applications cannot escape codon (Kurland, 1991) and mutational biases, they merely tilt the codon usage/mutation type of the organism rather than the innate “hard-coded” probabilities of specific codons mutating to those of amino acids. The predominance of specific tRNAs in species codon biases does not

intrinsically change the probability of Glycine codons (GGG, GGA, GGT, GGC) to mutate to Methionine (ATG). Rather, the possible biases that could affect the intrinsic mutation probabilities would arise in the form of misincorporation of specific base pairs during replication/transcription. Factors for such misincorporation can be contributed by the varying availability of specific bases in the cell or biases elicited by enzymes e.g. Cysteine deaminases that lead to an increased misincorporation of U(s) in the presence of specific drugs (Goulian *et al.*, 1980). Such biases, when analysed using the probability tables, allows an insight into the type of biological effects likely to be created.

CONCLUSION

We found statistical evidence that showed self-preservation and biases towards various amino acids in the genetic code table. In the event of substitution mutations, the highest probabilities are still conservative to steer away from aromatic, negative amino acids, and both start and stop codons. Such findings demonstrate self-preservation at the amino acid level occurring at the nucleotide level.

ACKNOWLEDGEMENTS

This research was funded by the Bioinformatics Institute core fund. We thank David Gunasegaran for his help in the calculations.

AUTHOR CONTRIBUTIONS

K.F.C. validated the manual calculation computationally. S.K. performed the calculations manually. S.K., K.F.C., J.Y.Y, W.S.D.K. and S.K.E.G. analysed the

results and wrote the manuscript. S.K.E.G. conceived and supervised the study. All authors read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Cao A, Galanello R. Beta-thalassemia. *Genetics in Medicine 2010;* 12**:** 61.

Chiang R Z-H, Gan S K-E, Su C T-T. A computational study for rational HIV-1 non-nucleoside reverse transcriptase inhibitor selection and the discovery of novel allosteric pockets for inhibitor design. *Bioscience reports 2018;* 38**:** BSR20171113.

Crick F H. Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology 1966;* 19**:** 548-555.

Crick F H. The origin of the genetic code. *Journal of Molecular Biology 1968;* 38**:** 367-379.

Crick F H, Barnett L, Brenner S, Watts-Tobin R J. General nature of the Genetic Code for Proteins. *Nature 1961;* 192**:** 1227-1232.

Freeland S J, Hurst L D. The genetic code is one in a million. *Journal of Molecular Evolution 1998;* 47**:** 238-248.

Goulian M, Bleile B, Tseng B. Methotrexate-induced misincorporation of uracil into DNA. *Proceedings of the National Academy of Sciences 1980;* 77**:** 1956-1960.

Haldane J B. The rate of spontaneous mutation of a human gene. *Journal of Genetics 1935;* 31**:** 317.

Hanahan D, Weinberg R A. Hallmarks of cancer: the next generation. *Cell 2011;* 144**:** 646-674.

Hollstein M, Sidransky D, Vogelstein B, Harris C C. p53 mutations in human cancers. *Science 1991;* 253**:** 49-53.

Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res 1984;* 12**:** 857-872.

Kurland C. Codon bias and gene expression. *FEBS Letters 1991;* 285**:** 165-169.

Lagerkvist U. " Two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences 1978;* 75**:** 1759-1762.

Ling W-L, Lua W-H, Poh J-J, Yeo J Y, Lane D P, Gan S K-E. Effect of VH–VL Families in Pertuzumab and Trastuzumab Recombinant Production, Her2 and FcγIIA Binding. *Frontiers in immunology 2018;* 9**:** 469.

Livingstone C D, Barton G J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Bioinformatics 1993;* 9**:** 745-756.

Roth D B, Craig N L. VDJ recombination: a transposase goes to work. *Cell 1998;* 94**:** 411-414.

Su C T-T, Ling W-L, Lua W-H, Haw Y-X, Gan S K-E. Structural analyses of 2015-updated drug-resistant mutations in HIV-1 protease: an implication of protease inhibitor cross-resistance. *BMC Bioinformatics 2016;* 17**:** 500.

Su C T-T, Ling W-L, Lua W-H, Poh J-J, Gan S K-E. The role of antibody Vκ framework 3 region towards antigen binding: effects on recombinant production and protein L binding. *Scientific reports 2017;* 7**:** 3766.

The International F.M.F. Consortium. Ancient Missense Mutations in a New Member of the RoRet Gene Family Are Likely to Cause Familial Mediterranean Fever. *Cell 1997;* 90**:** 797-807.

Tsui L-C. The spectrum of cystic fibrosis mutations. *Trends in Genetics 1992;* 8**:** 392-398.

Weiss R A. How does HIV cause AIDS? *Science 1993;* 260**:** 1273-1279.

FIGURES

a.a	No.of codons	T	A	C	G	T/A/C/G*
Phe	2	.104		.021		.031
Leu	6	.188	.042	.104	.042	.094
Ile	3	.083	.083	.021		.047
Met	1	.021	.021		.021	.016
Val	4	.104	.021	.021	.104	.063
Ser	6	.125	.063	.125	.063	.094
Pro	4	.021	.021	.188	.021	.063
Thr	4	.021	.104	.104	.021	.063
Ala	4	.021	.021	.104	.104	.063
Tyr	2	.063	.042	.021		.031
His	2	.021	.042	.063		.031
Gln	2		.063	.042	.021	.031
Asn	2	.021	.083	.021		.031
Lys	2		.104		.021	.031
Asp	2	.021	.042	.021	.042	.031
Glu	2		.063		.063	.031
Gly	4	.021	.021	.021	.188	.063
Arg	6	.021	.083	.104	.167	.094
Trp	1	.021			.042	.016
Cys	2	.063		.021	.042	.031
Stop	3	.063	.083		.042	.047

Figure 1. Base probability of amino acid in the genetic code. Colour scheme: Blue to red based on increasing probability.

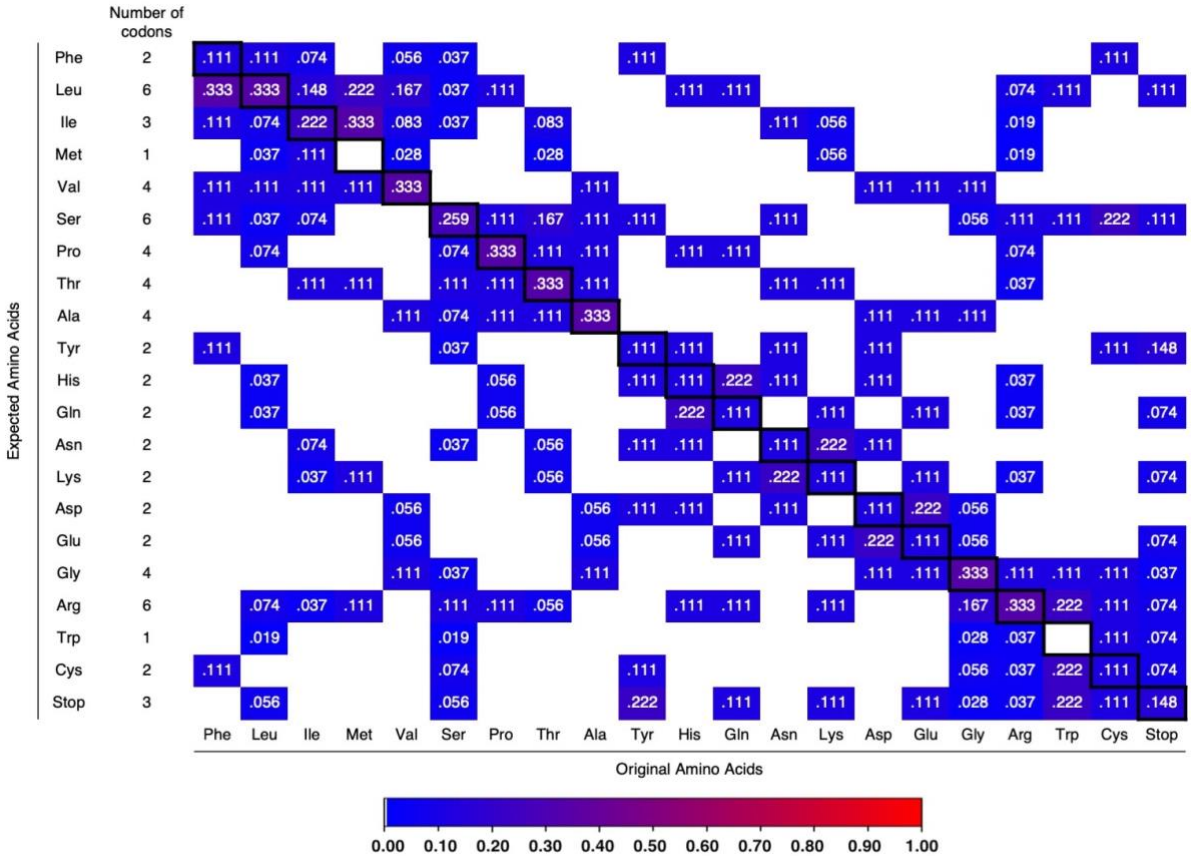

Figure 2. Probability of amino acids/stop codons changes. Colour scheme: Blue to red based on increasing probability. Probability tables of the amino acid change due to A, G, C, T mutations are separately calculated and shown in Supplementary Materials.

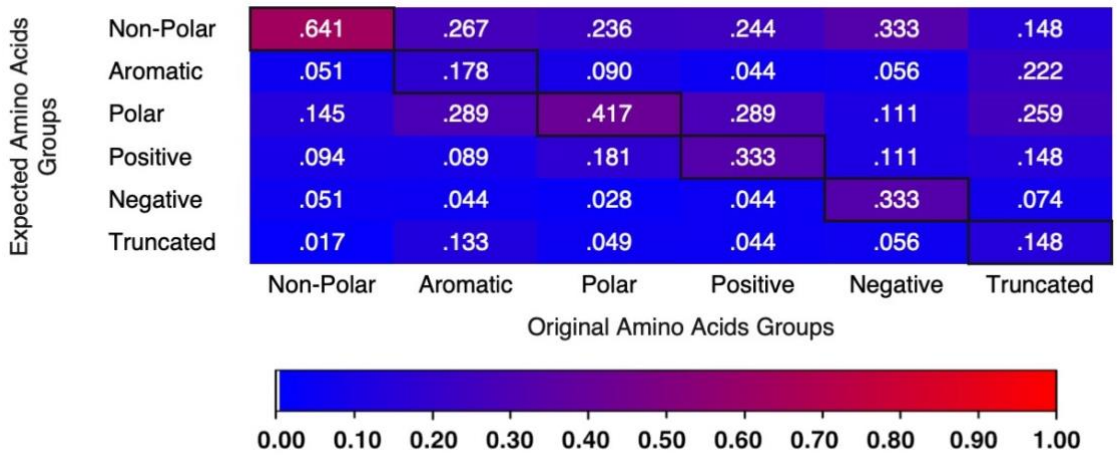

Figure 3. Probability of Physicochemical property changes (Livingstone & Barton, 1993). Colour scheme: Blue to red based on increasing probability.