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

* Corresponding Author: Tel: +65 6407 0584;

Email: samuelg@bii.a-star.edu.sg

† Both authors contributed equally to this work

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

SUMMARY

Mutations underpin the process of life[1], be it beneficial or detrimental. While mutations are assumed to be random in the lack of selection pressures[2], the genetic code underlies computable probabilities in amino acid phenotypic changes. Certain codons were found to require less mutational events to lead to specific amino acids than others. With applications in drug resistance[3,4], studying the probability of amino acid changes are 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 allow us to uncover an in-built self-preservation organization of the genetic code that steers away from disruptive changes at the amino acid level, away from start, aromatic, negative charged amino acids and stop codons. Our findings here provide the baseline mutational probability to study the genetic code mutations.

INTRODUCTION

Protein translation process DNA in the frame of three bases at a time referred to as a codon[5]. The open reading frame (ORF) typically begins with a Kozak sequence[6] embedding the start codon (ATG, coding for Methionine) and ends with a stop codon (TAA, TAG or TGA). The codons are then translated into amino acids based on the genetic code[7]. The genetic code has been described to be degenerative, where the majority of the common 20 amino acids encoded by multiple codons[5,8] and a degeneracy at the third base, called the Wobble Hypothesis [9].

Mutations in DNA underpin many life processes, and can occur as insertion, deletion and single nucleotide substitutions (SNS). Occurring at varying rates in daily life processes that range from hypermutations in the immune system[10], disease development in cancer[11,12], and drug resistance[3,4], they are essential for change.

Given that insertions and deletions cause frameshifts that are often detrimental in most cases, SNS are of interest, and they are further categorized into missense, nonsense and silent. SNS can also lead to disease states in classical examples that include the mutation in the β-globin gene from Glutamate (GAG) into Valine (GTG) as observed in patients suffering from sickle cell anaemia[13], cystic fibrosis[14] and beta thalassemia[15], where in the latter, the introduction of a premature stop codon truncates the polypeptide to result in a loss of function.

Analysing the genetic code table, we found clear biases to what changes mutations can achieve in limited mutation events. It is virtually impossible for ATG (Met) 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) within one such mutational event. Such limitations show 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.

RESULTS

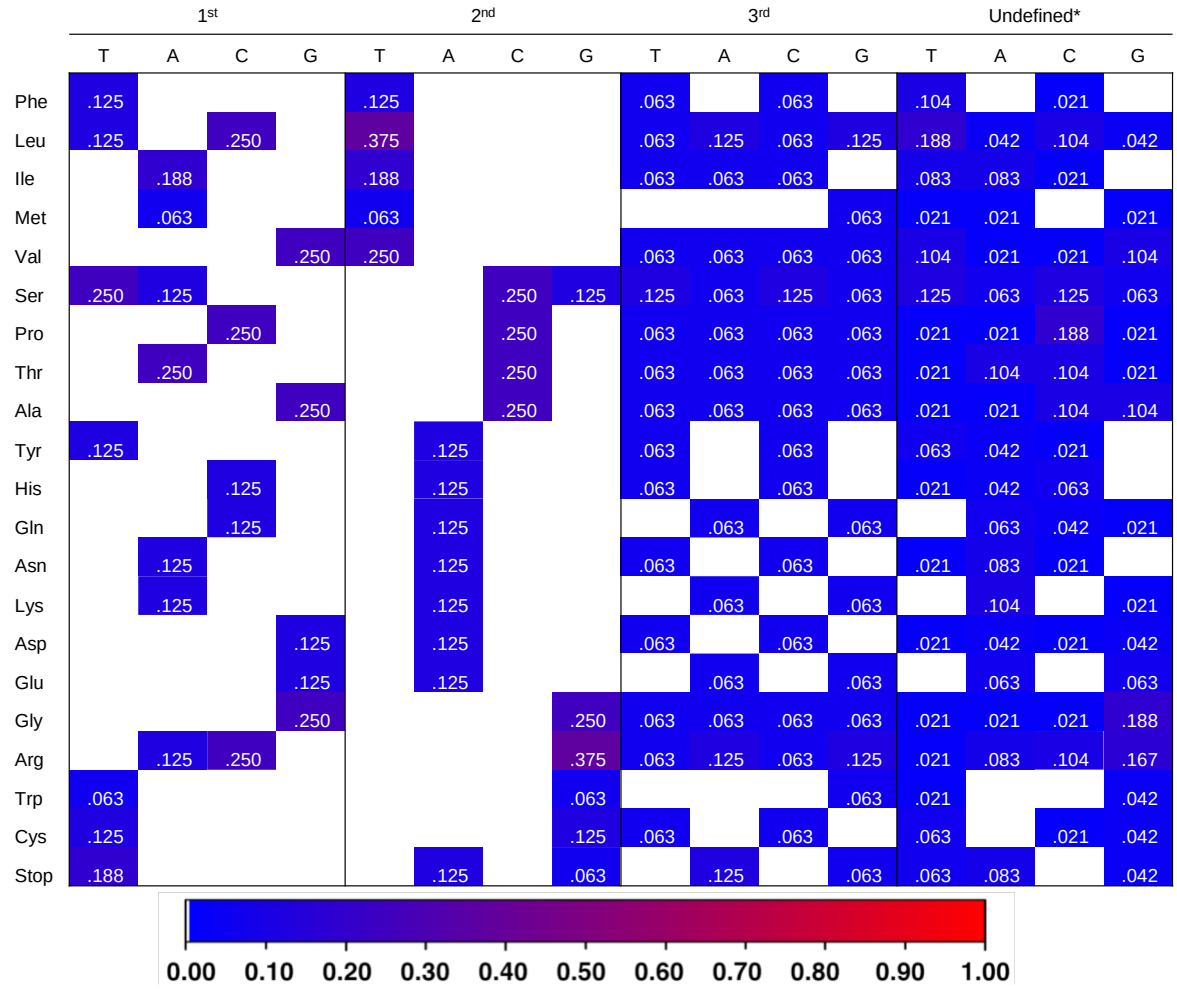

Figure 1. Probability of the corresponding amino acid (a.a.) occurring when the four bases are at the varying positions of the codon. Further calculations are in Extended Data Figure 1. Colour scheme: Blue to red based on increasing probability with white being p=0.

In analysing the probability of the 20 amino acids and stop codons when any of the codon bases are any of the four bases, there are clear biased predispositions towards certain amino acids (Figure 1, see Extended Data Figure 1A – C). For example, when having T in the first position, there is a bias towards codons encoding Serine (p=0.25) over other amino acids (p<0.2).

a.a	No. of Codons	T	A	C	G	T/A/C/G*
Phe	2	.190		.016		.052
Leu	6	.190	.032	.095	.032	.087
Ile	3	.079	.079	.016		.044
Met	1	.016	.016		.016	.012
Val	4	.095	.016	.016	.095	.056
Ser	6	.111	.048	.111	.048	.080
Pro	4	.016	.016	.286	.016	.084
Thr	4	.016	.095	.095	.016	.056
Ala	4	.016	.016	.095	.095	.056
Tyr	2	.063	.032	.016		.028
His	2	.016	.032	.063		.028
Gln	2		.063	.032	.016	.028
Asn	2	.016	.095	.016		.032
Lys	2		.190		.016	.052
Asp	2	.016	.032	.016	.032	.024
Glu	2		.063		.063	.032
Gly	4	.016	.016	.016	.286	.084
Arg	6	.016	.079	.095	.159	.087
Trp	1	.016			.048	.016
Cys	2	.063		.016	.032	.028
Stop	3	.048	.079		.032	.040

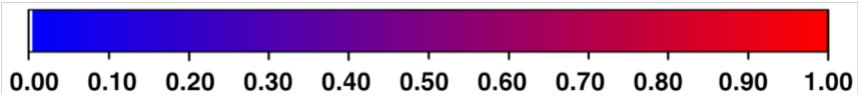

Figure 2: Probability of the SNS mutations (for both pre-determined or any of the T/A/C/G) at any position of the codon to give rise to the corresponding amino acid (a.a.) Colour scheme: Blue to red based on increasing probability.

Studying the effects of SNS on the 64 codons, we found that substitutions of T, A, C, and G into any location of the codon predisposed changes to Leucine/Phenylalanine, Lysine, Proline and Glycine, respectively (Figure 2). This is due to the dominance of specific bases in the respective codon makeup (Extended Data Table 1) of each amino acid. Amino acids such as Serine, with a more equal usage of the four bases, would demonstrate more balanced probabilities while those with a heavy bias towards a particular base, such as Phenylalanine (towards T) would be more dominant when mutational events involved the specific base (e.g. T).

We found some variations of probabilities between amino acids that have the same number of codons encoding them. While Serine and Arginine/Leucine have the same number of codons, they have distinct probabilities. On the other hand, other amino acids with four codons (Proline utilizing CC_ and Glycine GG_) have a higher probability compared to Serine with 6 codons (utilizing AG_ and TC_). The probability of the amino acids is thus also determined by the homogeneity of bases within the codons.

Statistically, biases towards T and A SNS can lead to 18 change possibilities (including the stop codon), whereas C and G SNS have only 16 change possibilities. (Figure 2). This is however balanced out when the various bases have equal possibilities when re-grouped into transition (A ↔ G) and transversion (C ↔ T) mutations.

Although both having only one codon, Methionine (ATG) is less probable to occur than Tryptophan (TTG). Similarly, the stop codons were less probable than other amino acids with the same number of codons, highlighting an intrinsic protection from mutations that that can lead to initiation and termination (Figure 2).

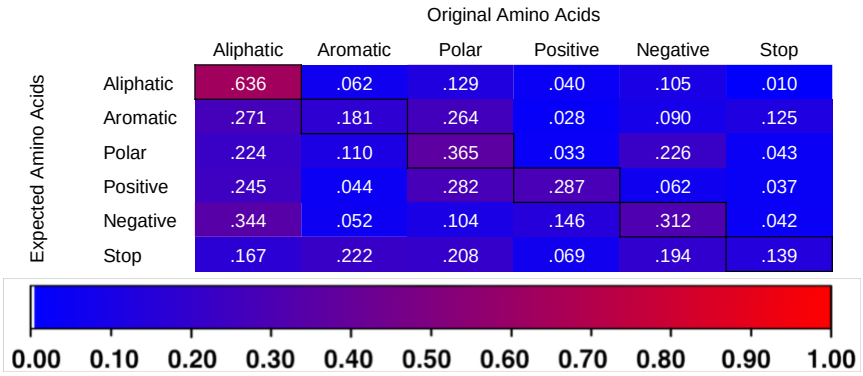

Figure 3. Probability of the SNS mutation (any of the T/A/C/G) at any position of the codon resulting in a change of amino acid properties[16]. Colour scheme: Blue to red based on increasing probability.

Probabilities of change are not uniform (Extended Data Figure 2A-D). Mutational biases to A have higher probabilities to become a stop codon, whereas a C SNS essentially steers away from stop codons. Unlike the other bases, A mutations are unlikely to lead to aliphatic amino acids and is predisposed towards polar neutral or polar positive amino acids. And T mutations are biased for aromatic amino acids while C mutations cannot lead to either start or stop codons.

The genetic code has an in-built self-bias to have high probabilities towards change within the same physicochemical amino acid group (Figure 3), except for aromatic, negatively charged amino acids, and the stop codon. The exceptions showed higher probabilities to change to the other group types. Interestingly, aromatic amino acids have the highest probability of change to a stop codon while aliphatic amino acids have an extremely low probability to change to a stop codon. This suggest that intrinsic barriers against certain drastic changes at the amino acid level present even at the nucleic acid level.

DISCUSSION

Upon analysis, the genetic code of life has in-built intrinsic biases and barriers. Within each mutational event, there are fixed probabilities for the type of change in selection-free conditions[17]. Our analyses have showed that C and G mutations encode less amino acids changes than A and T mutations, and in the latter mutations, encoding a broader range of amino acid changes (Figure 2).

Within the finite mutational events that can be incorporated into a population[18], the rate of diverse disruptive changes are further constrained by the mutational bias towards no change or a self-group at amino acid level (Figure 3) towards aliphatic amino acid codons. Such predispositions, even at the nucleotide level, supports the proposal of rarity of the genetic code[19] in the universe, especially given the organisation of self-bias to retain overall function at protein level.

While real-life applications cannot escape codon and mutational biases, they merely tilt the codon usage/mutation type of the organism rather than the actual probabilities of specific codons mutating to those of amino acids. The predominance of specific tRNA copy numbers in species codon bias do 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 due to the availability of specific nucleotides, or in the presence of deamination enzymes e.g. Cysteine deaminases that lead to an increased misincorporation of U(s) in the presence of specific drugs[20]. 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 predisposition towards amino acids in the genetic code table. In the event of mutations, the highest probabilities are 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.

REFERENCES

1 Hershberg, R. Mutation—the engine of evolution: studying mutation and its role in the evolution of bacteria. *Cold Spring Harbor perspectives in biology* **7**, a018077 (2015).

2 Lederberg, J. & Lederberg, E. M. Replica plating and indirect selection of bacterial mutants. *Journal of bacteriology* **63**, 399 (1952).

3 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* **17**, 500 (2016).

4 Weiss, R. A. How does HIV cause AIDS? *Science* **260**, 1273-1279 (1993).

5 Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the Genetic Code for Proteins. *Nature* **192**, 1227-1232, doi:10.1038/1921227a0 (1961).

6 Kozak, M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res* **12**, 857-872, doi:10.1093/nar/12.2.857 (1984).

7 Crick, F. H. C. The origin of the genetic code. *Journal of Molecular Biology* **38**, 367-379, doi:https://doi.org/10.1016/0022-2836(68)90392-6 (1968).

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

9 Crick, F. H. Codon—anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology* **19**, 548-555, doi:https://doi.org/10.1016/S0022-2836(66)80022-0 (1966).

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

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

12 Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253**, 49-53 (1991).

13 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* **90**, 797-807, doi:https://doi.org/10.1016/S0092-8674(00)80539-5 (1997).

14 Tsui, L.-C. The spectrum of cystic fibrosis mutations. *Trends in Genetics* **8**, 392-398 (1992).

15 Cao, A. & Galanello, R. Beta-thalassemia. *Genetics in Medicine* **12**, 61 (2010).

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

17 Kurland, C. Codon bias and gene expression. *FEBS Letters* **285**, 165-169 (1991).

18 Haldane, J. B. The rate of spontaneous mutation of a human gene. *Journal of Genetics* **31**, 317 (1935).

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

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

END NOTES

Acknowledgements

This research was funded by the Bioinformatics Institute core fund.

Author Contributions

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

AUTHOR INFORMATION

Competing Interests

The authors declare no competing financial interests.

Corresponding author

Correspondence and requests for materials should be addressed to samuelg@bii.a-star.edu.sg