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

Kwok-Fong Chan[1,†], Stelios Koukouravas[1,†], Joshua Yi Yeo[1], Wen-Shuo Darius 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 (S.K.E.G.)

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 applications 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 amino acid level. These changes include changes to start, aromatic, polar, 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 Met) 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 genetic code is degenerative at the third base, where the majority of the common 20 amino acids are encoded by multiple codons (Crick *et al.*, 1961; Lagerkvist, 1978). This degeneracy is 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 ATG (Methionine) 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.

RESULTS

Predispositions are present when analysing the probability of the 20 amino acids and stop codons when any of the codon bases are substituted with any of the four bases, (Figure 1, see Supplementary Figure S1A–C). Having a T in the first position biases towards codons encoding Serine ($p = 0.25$) over other amino acids ($p < 0.2$).

Studying the effects of SNS on the 64 codons, 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 (Supplementary Table S2). Amino acids such as Serine, with a more balanced usage of the four bases, have more distributed probabilities. Alternatively, amino acid codons with a heavy bias towards specific bases, such as Phenylalanine (towards T), would be dominant in the specific mutation type (e.g. T).

We found varying probabilities in amino acids that have the same number of codons. Serine and Arginine/Leucine have the same number of codons but have distinct

calculated probabilities. On the other hand, other amino acids with four codons (e.g. Proline utilizing CCn and Glycine GGn), have higher probabilities compared to Serine with 6 codons (utilizing AGn and TCn). Thus, the amino acid probability is determined by more than just the number of codons, but also by the homogeneity of bases within their codons.

Statistically, mutations towards T and A have 18 change possibilities (including the stop codon), whereas C and G substitutions have only 16 change possibilities (Figure 2). When re-grouped into transitions (A ↔ G) and transversions (C ↔ T) mutations, the various base mutations have equal possibilities.

Despite both being encoded by only one codon, Methionine (ATG) is less likely than Tryptophan (TTG) to occur. Similarly, the stop codons had lower probabilities than other amino acids with the same number of codons, highlighting intrinsic probability barriers for mutations to lead to initiations and terminations (Figure 2).

Probabilities of change to other amino acids were also not uniform (Supplementary Figure S3A–D and S4A–D). Mutational biases to A have higher probabilities to become stop codon and are unlikely to lead to aliphatic amino acids, being predisposed towards polar neutral or polar positive amino acids. T mutations, on the other hand, are biased towards aromatic amino acids. C mutations avert both start or stop codons (Supplementary Figure S3C), with a strong bias for Proline, and G mutations have strong pronounced biases towards Glycine and Arginine.

Calculating the probabilities of the 20 amino acids and stop codon to mutate to one another, there were interesting patterns observed when calculating the individual base changes (Supplementary Figure S3A-D). For example, in single G substitutions (Supplementary Figure S3D), Glycine and Tryptophan will only mutate to itself. Such limitations are also observed for Phenylalanine in T substitutions (Supplementary

Figure S3A), Asparagine and Lysine in A substitutions, and Proline in C substitutions. When studying all 4 base SNS together, there is a pattern of self-preservation shown in the diagonal line of Figure 3 having the highest probabilities in no amino acid change. The only exceptions to this are the single codon encoded Methionine and Tryptophan amino acids. Amongst all these, Glycine and Proline stand out in having highest probabilities of self-preservation.

When the amino acids are grouped into their physicochemical types, there are also self-biases to towards silent and conservative changes within the same amino acid group (Figure 3 and 4). However, this self-bias is absent for aromatic, polar amino acids, negatively charged amino acids, Histidine, and the start and stop codons. These exceptions had higher probabilities for nonsense and non-conservative change to the other physicochemical group types. After the highest probability of stop codon self-preservation, aromatic amino acids (especially Tryptophan and Tyrosine) have the next highest probability of change to a stop codon at $p = 0.125$ (Figure 4). Despite its large number of codons, aliphatic amino acids have extremely low probabilities to change to a stop codon at $p = 0.01$. 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 each mutational event, 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 than A and T mutations (Figure 2).

In the analysis of change of amino acids and stop codons to one another, there are strong self-biases in amino acid preservations with the exceptions of Methionine and Tryptophan which cannot be mutated given their single codon background (Figure 3). The high probabilities of Glycine and Proline towards self-preservation is highly interesting as there is a clear structural basis for this resistance to change. Glycine confers flexibility due to its small size, and Proline is known to have conformational rigidity. Thus, any mutations involving Glycine or Proline to other amino acids can significantly affect protein structure (Jacob *et al.*, 1999). Such in-built probabilities show an enigmatic preservation of protein structure at the genetic code level.

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 or within the physicochemical group (Figure 4) towards aliphatic amino acid codons. 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 tRNA copy numbers in species codon biases 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. Factors for such misincorporation can be 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 change 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.

Jacob J, Duclohier H, Cafiso D S. The role of Proline and Glycine in determining the backbone flexibility of a channel-forming peptide. *Biophysical Journal 1999;* 76**:** 1367-1376.

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

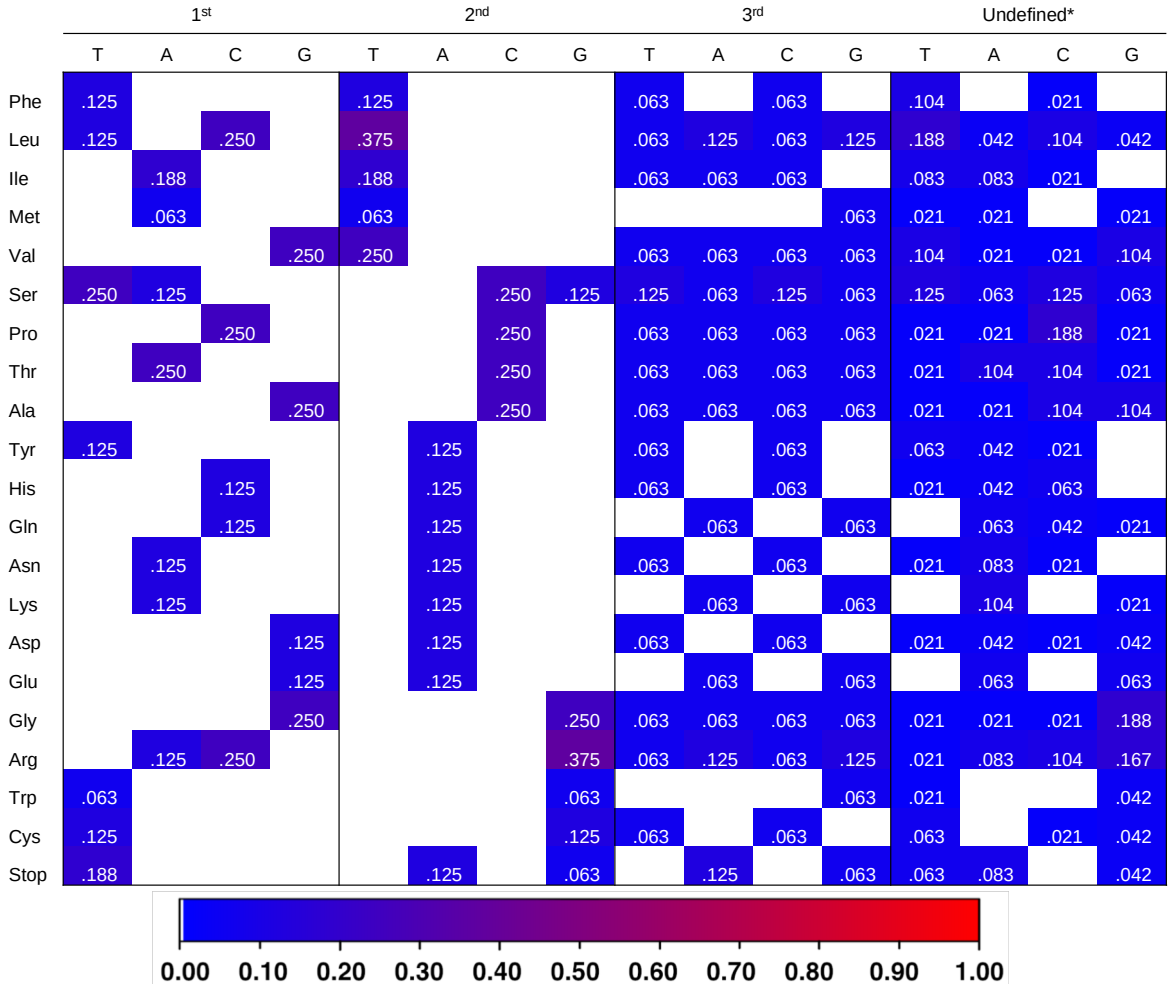

Figure 1. Probability of the corresponding amino acid (a.a.) occurring when the four bases are at the varying and undefined* positions of the codon. Further calculations are in Supplementary Figure S1. Colour scheme: Blue to red based on increasing probability with white being p=0.

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

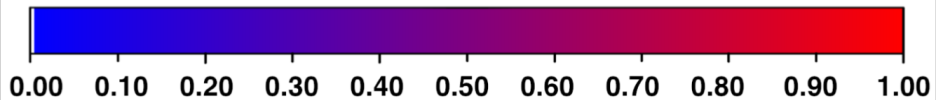

Figure 2. Probability of the SNS mutations (for both pre-determined or any of the T/A/C/G*) at any position of the codon to give rise to the corresponding amino acid (a.a.) Colour scheme: Blue to red based on increasing probability.

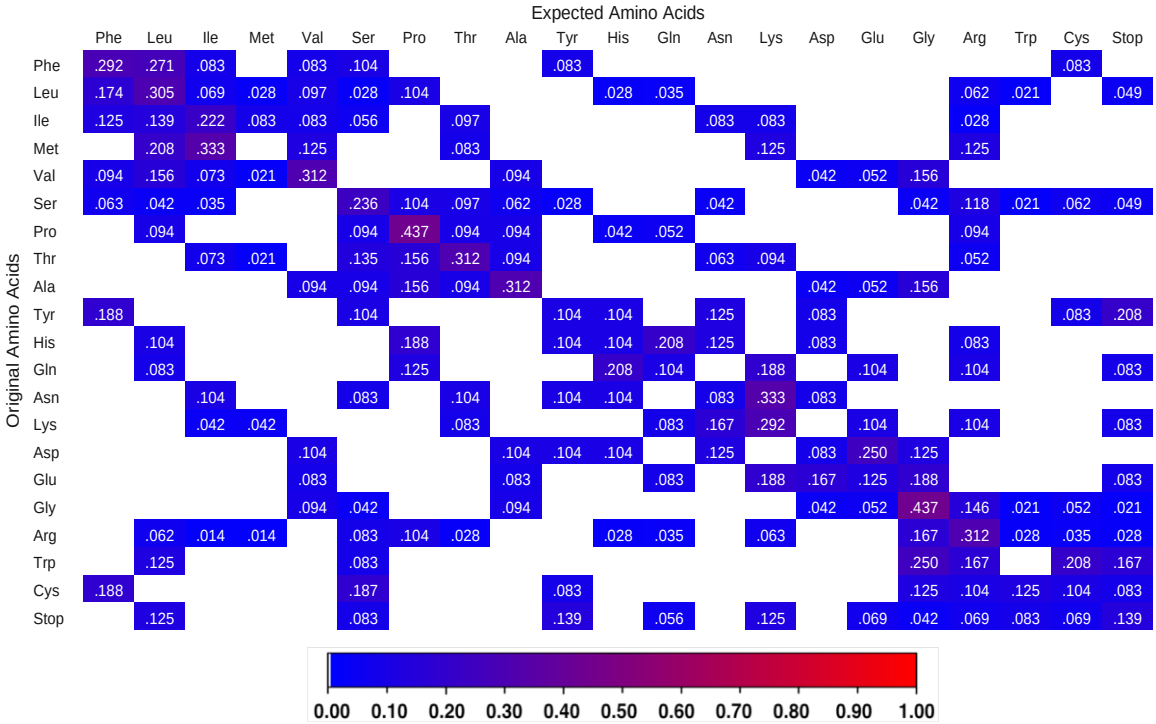

Figure 3. Probability of the SNS mutation (any of the T/A/C/G) at any position of the codon resulting in the amino acids. Colour scheme: Blue to red based on increasing probability. Probability tables of the amino acid change due to A, G, C, T mutations are separately calculated and shown in Supplementary Materials.

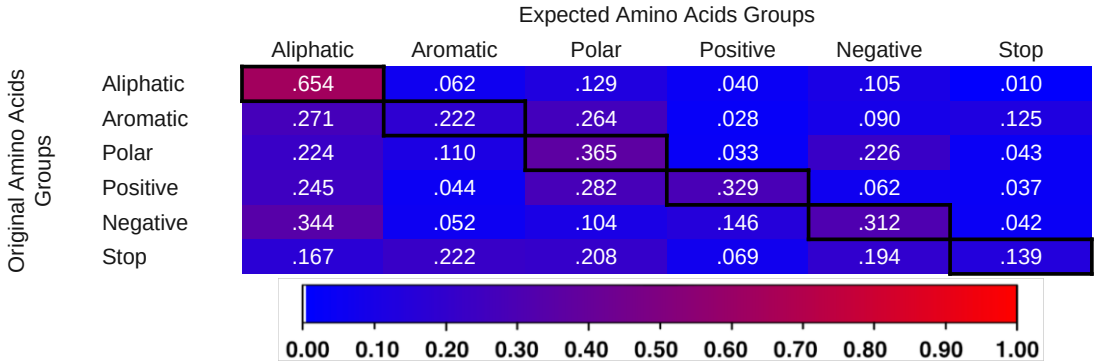

Figure 4. Probability of the SNS mutation (any of the T/A/C/G) at any position of the codon resulting in amino acids grouped by side chains(Livingstone & Barton, 1993). Colour scheme: Blue to red based on increasing probability.