RNA codons and correlant Amino Acids

Byoung-In Min

December 1st - 10th 2015

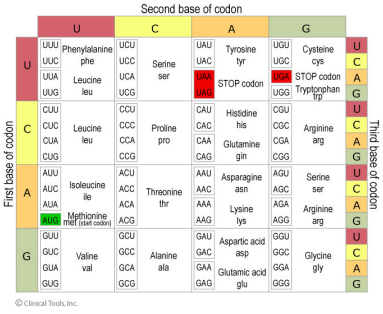
Abstract

Academics throughout the ages have attempted to describe the Universe using the language of Mathematics. From the movement of distant galaxies and rocket trajectories to the credit cards and computers we use every day, Mathematics pervades all aspects of life. It follows therefore, that mathematics may be used to describe DNA and RNA; molecules fundamental to life itself.

1 A brief insight to DNA sequencing

1.1 RNA structure

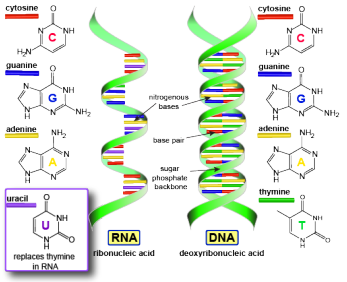
RNA codon table.jpg



DNA codes for every physical attribute we possess. For example, half of your code was in the head of a sperm which fused with your other half of your code in the centre of an egg to form a single cell with a single code - the zygote. This single cell then multiplied - along with its code - through the process of mitosis to form the trillions of cells and hundreds of different cell types found in an adult human being. This is accomplished in the most basic sense through subtle modifications in the reading of this single original code, and continues to function to dictate the daily workings of our bodies. DNA, like any code, must be read and processed to have any appreciable meaning. This takes the form of several enzymes. DNA polymerase 3 specifically reads the code and translates it into a complementary copy of DNA's functional counterpart, RNA[1]. This RNA code is further translated into proteins which lie behind most reactions that allows our cells, and by extension our bodies, to function. RNA and DNA are similar in structure, both are polymers composed of nucleotides. Each nucleotide consists of a phosphate group, a nitrogenous base and a five carbon sugar. DNA is double stranded and has deoxyribose as a sugar; single stranded RNA has ribose sugar and a slightly different complement of amino acids[2].

1.2 What are Codons?

RNA nucleotides have a complement of 4 basic nitrogenous bases, A, C, G and U, standing for Adenine, Cytosine, Guanine, and Uracil. These form base pairs with each other where A always binds to U and C to G. The reason for this is so that the width of a DNA strand can be consistent throughout. DNA shares three of these bases, however contains Thymine as a substitute for Uracil[2].



1.3 The Triplet code

Nucleotides are arranged in triplets, each triplet corresponding to either one of twenty amino acids or a STOP codon. It is a very efficient way to code for the essential 20 amino acids needed for our bodies. If the sequence was in duplets, the total combinations would be

$$4^2 = 16 \tag{1}$$

This obviously is not sufficient to code for 20 unique amino acids. As Triplets however,

$$4^3 = 64 \tag{2}$$

there are plenty of codons available to code for the amino acids.

2 Frequency of codons per amino acid

2.1 Distribution of Codons

Assuming that each base had an equal probability of being chosen (i.e. 1/4), from 1.3 it is evident that each amino acid will have more than 1 codon to code for itself. If the distribution of codons to amino acids were uniform, then each amino acid would have;

$$64/20 = 3.2 \text{ codons}$$

However, as we can see from the table in 1.1, this is not the case at all.

2.2 Frequency of each amino acid

This naturally raises the question, in a randomly generated sample, how many amino acids will be leucine, proleine or any amino acid in question? By the table from 1.1, we could hypothesise that it would be easily worked out via

$$\frac{number of codons for amino acid in question}{total number of codons} \times 100$$

To test this, we will assume that the probability of each nucleotide being present to all be a quarter. In reality, the probability per base varies considerably depending on the relative position and function of the sequence but will be generalised here for simplicity. In humans the probability is around 0.3 for A, 0.3 for U, 0.2 for G and 0.2 for C [5].

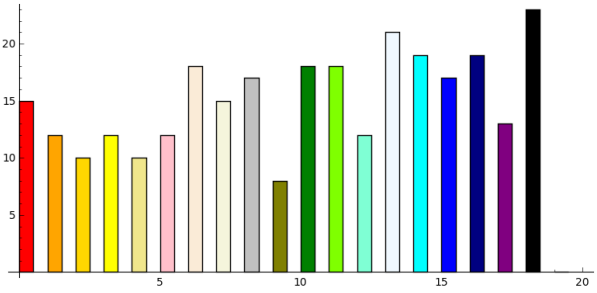
$$P(A) = P(U) = P(C) = P(G) = 0.25 \text{ codons}$$

A random sample of a 1000 codons will be printed and the correspondent amino acids will be counted by the following code.

```
#definition of code that generates 3 size code.
def id_generator(size=3, chars='AUCG'):
    return ''.join(random.choice(chars) for _ in range(size))

#variable counters
leucineCount=0#CUU CUC CUA UUA UUG

#loop of creating 1000 codons
for x in range(0,1000):
    #generate id
    codon1 = id_generator()
    if codon1 == ('CUU' or 'CUC' or 'CUA' or 'UUA' or 'UUG'):
        leucineCount = leucineCount+1

        print ('within 1000 iterations:')
print ('number of leucine: '+ str(leucineCount))
```

The rest of the code can be found at [4]

Thus, although the numbers per amino acids range between 8-24, it is evident that each amino acid has a distinct probability and thus cannot be modelled by uniform. From this graph, we can see that amino acids with more possible codons have a higher frequency, whilst those with smaller possible codons have a lower frequency.



Note, the start codon and stop codon are not printed. Stop codons do not have a correlant amino acid whilst the start codon AUG does translate to an amino acid called methionine, but the methionine coded by the start codon AUG often cleaved from the sequence so the frequency of methionine will be less than in reality.

3 Mean length of DNA fragment caused by a restriction enzyme

Theoretically speaking, DNA could be coded infinitely. The result of course is only imaginable, but the question this problem derives is; at what average length does a stop codon actually appear? Hypothetically, it could be straight away, or infinity. The following code prints random lists of codons starting with AUG and ending with a stop codon. It counts the number of codons in between the start and stop codon.

```
#looping combinations until the codon is equal to 'AUG'
def codoncounting():
    #initialising boolean variables
    loop1 = 0
    loop2 = 0
    #variables
    codoncount= 0
    while loop1 == 0:
        #generates the id
        codon2 = id_generator()
```

The rest of the code can be found [4] .
We can see that the average length is around 20. Indeed, this corroborates our findings in 1.1 as there are 3 stop codons. If we calculate the number of stop codons over the total number of stop codons multiplied by the average length it is approximately equal to the total probability.

$\frac{3}{64} \times averagelength \approx 1$

4 ABCC11 Protein

Naturally, it seems odd that these series of letters have a significance to our lives. To respresent the significance of DNA, I will use an example phenotype where only a single gene is involved, the gene coding for the ABCC11 protein. [3]

4.1 Phenotype

	Ile	Ala	Ser	Val	Leu	Gly	Pro	Ile	Leu	Ile	Ile	Pro
wet earwax	ATT	GCC	AGT	GTA	CTC	GGG	CCA	ATA	TTG	ATT	ATA	CCA
dry earwax	ATT	GCC	AGT	GTA	CTC	AGG	CCA	ATA	TTG	ATT	ATA	CCA
						Arg						

Bases 523-558 of the coding sequence of *ATCC11*, along with the amino acid sequence. The DNA polymorphism at site 538 causes the amino acid polymorphism that determines earwax type.

earwax.png

The protein ABCC11 codes for the consistency of earwax and has two forms, wet and dry. The majority of people have wet earwax, as this gene is dominant over dry earwax. They differ in a single nitrogenous base change in the DNA sequence, yet result in a completely different phenotypic effect. This demonstrates the power of this simplistic code in all aspects of our physical makeup.

4.2 Phenotype

The more interesting element of this single nucleotide difference is how agreeable your sweat is to bacteria. Those with the AAA codon will have body odour as the bacteria is more attracted to that moisture and causes the unpleasant scent. However, those with GAA produce sweat which is less habitable to bacteria[3].

5 Conclusion

All organisms have a genetic code. With each essential amino acid having a unique code but altering probabilities, it is one of the core reasons for life to exist in such diversity. Indeed, although this code has many assumptions limited in its choice and variety, it only proves to us how unique each and every organism is including ourselves. Also, to see how 1 nucleotide difference can produce dramatic changes within an individual demonstrates the importance of this ubiquitous molecule.

References

[1] Byoung. Dna picture on dna replication.

[2] Neil A. Campbell and Jane B. Reece. *Biology*. Pearson Education, 2005.

[3] John McDonald. Myths of human genetics.

[4] Byoung-In Min. Code for cousework and graph.

[5] ugrad. Applications of probability.

1.Link to my code; https://cloud.sagemath.com/projects/17e59559-a6ad-4ed9-ba58-4c95144c3a8f/files/Computing
2.Link to my bibliography;
https://cloud.sagemath.com/projects/17e59559-a6ad-4ed9-ba58-4c95144c3a8f/files/bibliography.bib