An Evaluation of the Quality of Codon Deviation Coefficient as a Measurement of Codon Usage Bias

Amaral Lab

Ali Saeed[1], Sophia Liu[2]

[1]Adlai E. Stevenson High School, Lincolnshire, IL

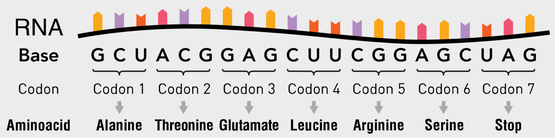
[2]Northwestern University, Dept. Chemical and Biological Engineering

Stevenson High School

Background

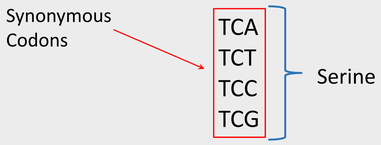
DNA is composed of four nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C), which form **nucleotides**.

Sequences of three nucleotides, called **codons**, are translated into amino acids. There are 64 possible codon combinations; the 61 of which that code for amino acids are called **sense codons**.



However, there are only 20 translated amino acids. Different codons which translate into the same amino acid are called **synonymous codons**.

Some synonymous codons are used with greater frequency than others, a phenomenon known as **codon usage bias (CUB)**.



Why is Codon Usage Bias a Big Deal?

- Increased translational efficiency and accuracy
- Provides evolutionary information
- Optimize genetic engineering

Codon Deviation Coefficient

Codon Deviation Coefficient (CDC) is a metric used to calculate CUB using adenine, thymine, guanine, and cytosine content *(A, T, G,* and *C),* **GC content** *(S),* and **purine content** *(R).*

Calculating CDC

1. Calculate *GC content (S)* and *purine content (R)* for each of the three codon positions *(i)*

$$S_i = \frac{G + C}{number\ of\ sense\ codons} \qquad R_i = \frac{A + G}{number\ of\ sense\ codons}$$

2. Calculate *expected nucleotide contents* (A_i, T_i, G_i, C_i) for each of the three codon positions

$$A_i = (1 - S_i)R_i, \qquad T_i = (1 - S_i)(1 - R_i)$$
$$G_i = S_i R_i, \qquad C_i = S_i(1 - R_i)$$

3. Calculate the *expected codon usage (π)* for each of the 61 sense codons

For example, using the codon AGC:

$$\pi_{AGC} = \frac{A_1 G_2 C_3}{\sum x_1 y_2 z_3}, where\ x, y, z \in \{A, T, G, C\}$$

4. Calculate the *observed codon usage* ($\check{\pi}$) for each of the 61 sense codons

For example, using the codon AGC:

$$\check{\pi}_{AGC} = \frac{number\ of\ times\ AGC\ appears\ in\ sequence}{number\ of\ codons\ in\ sequence}$$

5. Calculate the CDC from the observed and expected codon usages of each of the 61 sense codons

$$CDC = 1 - \frac{\sum_{xyz} \pi_{xyz} \times \check{\pi}_{xyz}}{\sqrt{\sum_{xyz} {\pi_{xyz}}^2 \times {\check{\pi}_{xyz}}^2}} \qquad \pi_{AGC} \approx .0096, \quad \check{\pi}_{AGC} = 1/6 \quad \ldots$$

Pos	Pos	Pos
1	2	3
A	G	C
G	G	A
A	A	A
C	C	A
G	C	G
G	C	G

i	S_i	R_i
1	2/3	5/6
2	5/6	1/2
3	1/2	5/6

i	A_i	T_i	G_i	C_i
1	5/18	1/18	5/9	1/9
2	1/12	1/12	5/12	5/12
3	5/12	1/12	5/12	1/12

$$CDC \approx 0.490$$

CDC values range from 0 (no codon usage bias) to 1 (maximum bias)

Goal

- Design tests to determine the quality of CDC as a metric for measuring CUB

Hypotheses

How would CDC behave if it were a robust metric?

1. CDC will converge to **zero** if the sequence reflects **no codon usage bias**
2. CDC will converge towards **one** if the sequence reflects **maximum bias**
3. CDC will be **independent of the length** of the sequence
4. CDC will indicate **no correlation to the GC content** of the sequence

Method

- Use Python to generate random sequences from predetermined codon probabilities
- Measure the CDC as more codons are appended to the sequence

Hypothesis #1

METHOD

How can we test the hypothesis?

- CDC values will be calculated along a randomly generated sequence in which each codon has an equal probability of appearance

$$P(x) = \begin{cases} 1/61, & x \in ATT \\ 1/61, & x \in ATC \\ 1/61, & x \in ATA \\ & \ldots \end{cases}$$

Hypothesis #2

METHOD

How can we test the hypothesis?

- CDC values will be calculated along a randomly generated sequence in which **only one random synonymous codon** per each set of 20 synonymous codon groups has an equal probability of occurrence

$$P(x) = \begin{cases} 0, & x \in ATT \\ 0.05, & x \in ATC \\ 0, & x \in ATA \\ & \ldots \end{cases} \rightarrow \text{Isoleucine}$$

Hypothesis #3

METHOD

How can we test the hypothesis?

- CDC values will be calculated for a sequence that consists of **repeated sequences of all 61 codons**
- CDC will be calculated at every repeat

AAA AAT AAC AAG...CCG CCC AAA AAT AAC AAG...

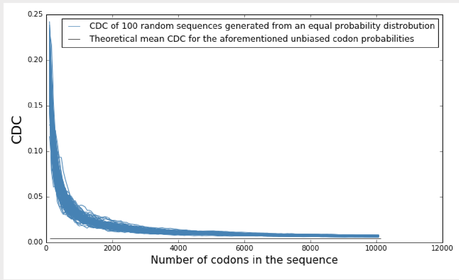
↑ Calculate CDC!

Hypothesis #4

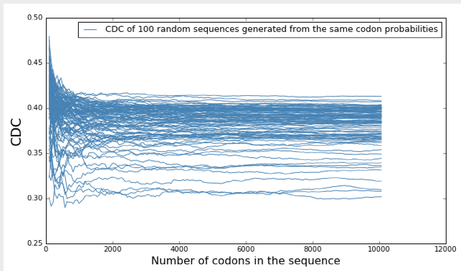
METHOD

How can we test the hypothesis?

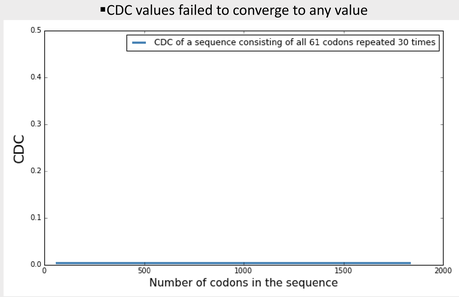
- CDC values will be calculated for completely unbiased sequences with equal amino acid usage but **variable GC contents**



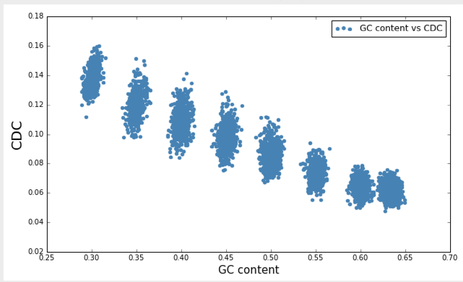
- CDC values converged towards **zero**, reflecting no bias



- CDC values failed to converge towards one and reflect the highly biased sequence
- CDC values failed to converge to any value



- CDC values remained constant after every repeat, showing no correlation to length
- CDC values did not exactly equal zero



- CDC values demonstrated an undesirable **correlation to the GC content**

Conclusions

❖ CDC accurately measures unbiased sequences

When measuring sequences in which every codon is used with equal probability, the CDC value should and does converge to zero.

❖ CDC shows no correlation to the length of the sequence

Codon usage bias values should not demonstrate correlation to sequence length, which is an extraneous quality.

❖ CDC fails to accurately and precisely measure highly biased sequences

The CDC metric fails to converge to one while measuring the most highly biased situation in which only one synonymous codon is assigned a probability of occurrence.

❖ CDC demonstrates an undesirable correlation to the GC contents of sequences

Codon usage bias values should not demonstrate correlation towards GC content, which is an extraneous quality.

Acknowledgements

I would like to thank Dr. Luis Amaral and the members of his lab at Northwestern for allowing me to conduct research and being friendly and helpful along the way. In particular, I sincerely thank graduate student Sophia Liu for coordinating my project and assisting me through every step.

References

Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. 33(4): 1141-1153

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. GENETICS. 136(3): 927-935