Geometric Structure of Codon Relationships

I wish to suggest a three-dimensional structure for relating codons to amino acids. It is also useful for relating codons to other codons. This structure has novel features that are of considerable biological interest.

The nearly universal correlation between codons and amino acids in genetic translation has already been described. The correlation data is seen as 'linear' or one-dimensional, because each codon is assigned to only one amino acid, but the data is commonly presented in a two-dimensional grid. I wish to put forward a radically different geometric model for codon assignments - The Rafiki Model - that arranges the data in a three-dimensional symbolic network.

'The Genetic Code' is in part a molecular substitution cipher, where codons are translated into amino acids. The substitution key has been known for almost four decades, but a logical origin of the cipher, and a shared meaning between sequences is still unknown. As a cipher, it is a molecular form of cryptography. In other words, meaning is encoded in one molecular sequence and decoded into another. However, it might also be an example of steganography, where the physical form of the message conceals a part of its meaning.

In this case, genetic information is stored in a geometric structure with essentially no variation - DNA's double helix - so nucleotides display a periodic spatial arrangement[1]. It is therefore assumed that no spatial information is directly decoded from nucleotides, and protein shapes are derived entirely post-translation. According to a one-dimensional paradigm, the final structures of proteins are pre-determined by epigenetic forces acting upon amino acid sequences alone. But this view fails to

account for the significant role played by other molecules, and there are tenable alternatives. It is possible that relationships between codons, in concert with consecutive tRNA interactions, can communicate spatial information directly into various conformations of peptide bonds. And beyond that, the proteome has a powerful voice in reading genetic messages.

Encoding and decoding any form of information involves the transfer of meaning from one perception to another, and in all cases it requires that both share an encryption key. Apart from the codon table itself, the linear cipher provides no clue about the shared molecular meaning between nucleotides and amino acids. It is therefore difficult to imagine how these molecular sets could develop, utilize and maintain a consistent key for encryption in one dimension alone. However, nucleotide triplets do share a cryptic geometry with amino acids, and this can provide a framework for crystal cryptography.

A dodecahedron is made of twelve pentagonal faces and twenty vertices. Each vertex is a collection of three faces, and from each we are able to generate six permutations of faces. All sixty-four codon permutations can be represented by a single dodecahedron with properly labelled faces (Figure 1). This dodecahedron also embodies a logical collection of one-hundred-and-twenty discrete internal tetrahedrons. Each permutation of faces on the dodecahedron is coincident with the symmetry of exactly one internal tetrahedron. Conveniently, the symmetry of the DNA double helix is that of sequential dodecahedrons, and proteins are sequences of tetrahedrons. The two sets of molecules can use this shared symmetry as the logical basis for an encryption key.

Viewed in this three-dimensional network, the assignment data reveals harmonic relationships between codons and several physical properties of amino acids, such as molecular size, charge, and most significantly, water affinity[2]. These complex relationships provide a systematic logic for molecular encryption.

Decoding messages in the proteome is a multi-dimensional dance of molecular computation, and the proven influence of 'synonymous' codon substitutions make 'silent' mutations anything but silent in the translation of genetic information.[3, 4, 5] Recent advances in the study of protein structures have shown that identical amino acid sequences do not always produce identical protein folds.[6, 7, 8, 9] Furthermore, structures at all scales are conserved far more than amino acid sequences.[10, 11, 12] The current body of knowledge regarding genetic translation makes a strict interpretation of the one-dimensional paradigm virtually untenable. In the broader scope of biology, the purview of genetic information starts in sequences of nucleotides, but extends well beyond sequences of amino acids.

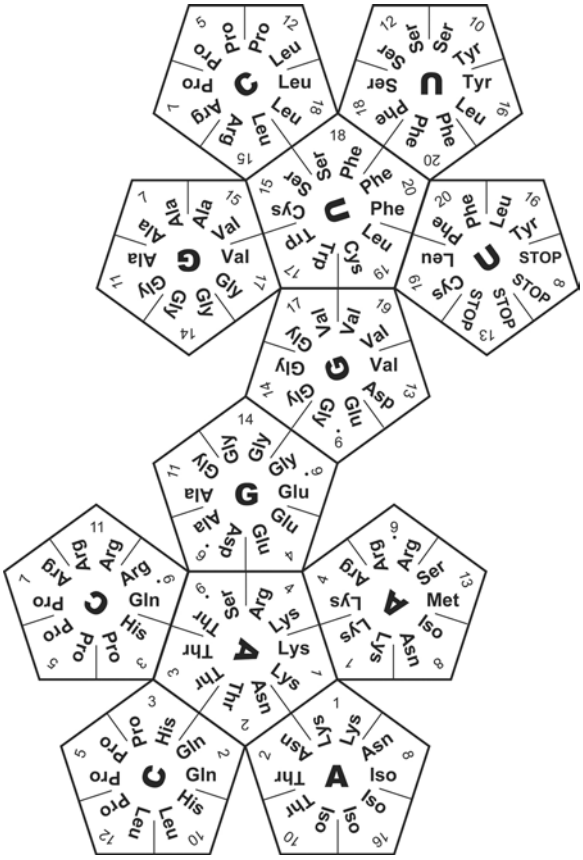
The Genetic Code is an Enigma*. Nucleotide messages are input and then appear as random amino acid ciphertext.[13] These messages are decoded and further processed in the proteome,[14] making basic studies of codon relationships eminently useful. Regardless of the mechanism, we know that nucleotide sequences impact genetic computations in many dimensions. Watson and Crick gave us the shape of the genetic hard drive, to our great advantage. Perhaps it is also to our benefit that we find the shape of the genetic processor.

Mark White, MD

Correspondence and requests for additional materials should be addressed to Mark White, email rafiki72@insightbb.com.

Figure 1.

Rafiki Model of codon assignments. The figure can be folded into a dodecahedron. Each vertex formed by the intersection of three faces can be permuted in six ways. The first amino acid in the proper reading direction (clockwise or counter-clockwise) represents the codon assignment.



---

[1] Watson, J. & Crick, F. Molecular Structure of Deoxypentose Nucleic Acids. *Nature* 737-738 (1953).

[2] Bashford, J. D. & Jarvis, P.D. The genetic code as a periodic table: algebraic aspects. http://www.physics.adelaide.edu.au/~jbashfor/ (2000).

[3] Deana, A., Ehrlich, R. & Reiss, C. Silent mutations in the Escherichia Coli ompA leader peptide region strongly affect transcription and translation in vivo. *Nucleic Acids Research*. Vol. 28, No. 20, 4778-4782 (1998).

[4] Carlini, D. & Stephan, W. In Vivo Introduction of Unpreferred Synonymous Codons Into the Drosophila Adh Gene Results in Reduced Levels of ADH Protein. *Genetics* **163**: 239-243 (2003).

[5] Humphreys, D. et al. High-Level Periplasmic Expression in Escherichia Coli Using Eukaryotic Signal Peptide: Importance of Codon Usage at the 5' End of the Coding Sequence. *Protein Expression and Purification* Vol. 20, No. 2, 252-264 (2000).

[6] Weiss, M., Jabs, A. & Hilgenfeld, R. Peptide bonds revisited. *Nature Structural Biology* Vol. 5. No. 8, 676 (1998).

[7] Oresic, M. & Shalloway, D. Specific Correlations between Relative Synonymous Codon Usage and Protein Secondary Structure. *Journal of Molecular Biology*, Vol. 281, No. 1, 31-48 (1998).

[8] Gupta, S., Majumdar, S., Bhattacharya, T., Ghosh, T. Studies on the Relationships between the Synonymous Codon Usage and Protein Secondary Structural Units. *Biochemical and Biophysical Research Communications* Vol. 269, No. 3, 692-696 (2000).

[9] Cortazzo, P., et al. Silent mutations affect in vivo protein folding in Escherichia Coli. Biochemical and Biophysical Research Communications Vol. 293, No. 1, 537-541 (2002).

[10] Denton, M. & Marshall, C. Laws of form Revisited. Nature Vol. 410, pg 417 (2001).

[11] Thorne, J. L. Models of protein sequence evolution and their applications. Current Opinion in Genetics & Development 10:602-605 (2000).

[12] Bastolla, U., et al. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *Journal of Molecular Evolution*. 56(3):243-54 (2003).

[*] Enigma was a German cryptographic machine in wide use during WWII. Polish and British cryptanalysts shortened the war and saved many lives by mastering the cipher logic of Enigma. Alan Turing, who is now recognized as a visionary of modern computers, made the most famous contribution.

[13] Nevill-Manning, C. & Witten, I. Protein is incompressible. http://craig.nevill-manning.com/publications/DCC99.pdf (2000).

[14] Tyers, M. & Mann, M. From genomics to proteomics. *Nature* Vol. 422, 193-197 (2003).