

Maximum Symmetry in the Genetic Code

The Rafiki Map

by Mark White, MD

Introduction

The genetic code, as it is presently defined, is merely a set of data. I wish to propose a novel structure for the data to maximize its symmetry. The physical structure of DNA, by analogy, is central to nature's methods of storing and replicating genetic information. The double helix has a precise form to dictate its function, which shows that relationships between molecules - not the molecules themselves - define the essence of biological information. These relationships are described by structure and shape. What then is the shape of the genetic code?

We conventionally view the genetic code as a two-dimensional spreadsheet, where nucleotide symbols standing for the first and last codon positions are arranged in rows, and the symbols for the middle nucleotides are arranged in columns.

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

This data structure is certainly familiar enough to obviate further description here. Of course all data structures must adopt conventions that are predicated on some form of subjectivity, so we must accept this in our symbolic treatments of the code. **However, the genetic code itself must be based on a purely objective structure in nature.** Imagine organic molecules given the task of obeying the dictates of the genetic code. What informative, objective parameter could possibly exist in the universe to serve as a logical foundation for this if not shape?

Any spreadsheet standing for the genetic code must employ a reading algorithm that permutes single nucleotides into triplets. The data cells spatially created by the permutation are then mapped with the twenty amino acids contained in the standard set. There are unfortunate, subtle biases to this particular schema, and they bring with them seldom-recognized yet nasty epistemic side effects. First, the inherent symmetry of nucleotide relationships is destroyed by unbalanced treatment of the three positions within permuted triplets. Second, a completely subjective hierarchy of nucleotides must be imposed on the data structure (commonly U, C, G, A is the one chosen) and this relative positioning of symbols impacts the visible patterns commonly recognized in the code. Many alternate structures have been proposed; they include hypercubes, circles, spirals and various funky geometric schemas, but they all share similar drawbacks due to a lack of maximum symmetry in the combined components of the code. At one time in the late 1960s, the framework of a 5,000 year-old Chinese spiritual philosophy known as the I Ching was formally and studiously fitted with the genetic code. No kidding. This demonstrates the importance of, and the extent to which we are willing to investigate and expand our perspective on possible structures for this particular set of data.

Regardless of the actual structure used in viewing the genetic code, **an unmistakable element of symmetry** usually emerges from the data. Symmetry in this case can be defined in many ways, but most of them relate to a transformation of data that leaves fundamental properties unchanged. For example, a wheel is transformed by rotation, but symmetry in the perimeter and spokes leave the appearance of the wheel unchanged. In the case of the genetic code, symmetries have been found throughout the data using a variety of creative techniques. Appreciating many of these symmetries in the past has required high-level, abstract mathematics, as well as taking account of several properties of the molecules in the genetic code. The most obvious symmetry is found in the third position of codons, where the same amino acid is often assigned independent of the nucleotide in that position. This symmetry in the data is associated with concepts known as 'degeneracy' of codon assignments and 'wobble' in cognate anticodons.

The goal of this paper is to describe a symbolic configuration of nucleotides that is maximally symmetric. I call this the Rafiki map. Since the data is based upon nucleotide permutations, it is logical that symmetry in the data should be as well. This means that all nucleotides must be treated

independent of position or relationship to other nucleotides within the structure. Remarkably, it can be achieved by using twelve faces of a dodecahedron, a schema that magically generates all sixty-four codons. This treatment of the data results in the most symmetric and therefore not surprisingly the most compressed possible arrangement of nucleotides. Through a spherical permutation network of single nucleotide symbols we also are able to find the most symmetric arrangement of nucleotide doublets and triplets. Preserving all of these natural symmetry elements in the overall data structure is informative to a proper view of the genetic code.

Constructing the Rafiki Map.

Constructing a codon map from a dodecahedron is conceptually quite simple. Begin with a tetrahedron, and label each of its four vertices with one of the nucleotides in mRNA.



These tetrahedral vertices serve as major poles in the codon map. An equilateral triangle labeled with three congruent symbols is then affixed at proper angles to each pole.

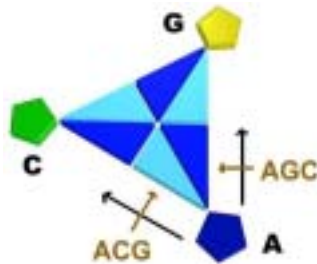


These four symbolic triads comprise the twelve faces of a dodecahedron, as well as the twelve vertices of an icosahedron. The tetrahedron, dodecahedron and icosahedron are three of only five possible regular solids, also known as platonic, or perfect solids. The other two perfect solids, cube and octahedron, can be applied to subsets of the data as well. The dodecahedron and icosahedron are dual to each other, a relationship where vertices in one solid

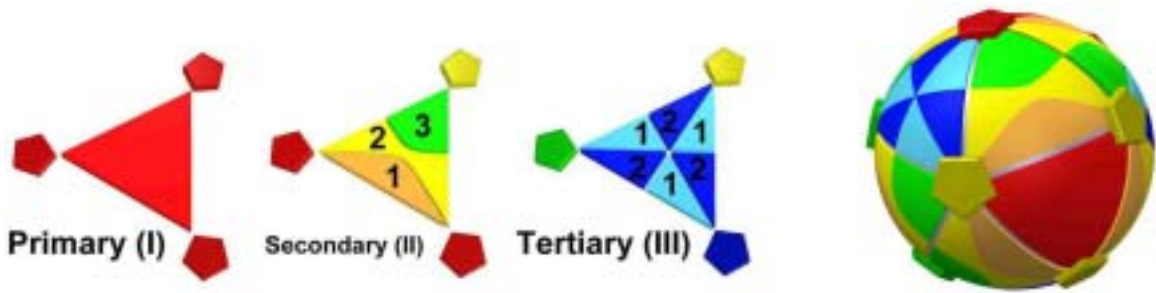
match face centers in the other. The cube and octahedron are duals also, but the tetrahedron is dual only with itself.



There are 120 distinct rotational permutations of three adjacent faces on a dodecahedron. Each permutation defines one of twenty vertices, or twenty distinct symbolic triangles. The codon reading conventions within face triplets are as follows. Begin with any face on the dodecahedron and move to any of the five adjacent faces. There are now only two remaining faces that are adjacent to both of the first two, and choosing one of them defines a vertex. Selecting three ordered faces in this way creates one of six permutations from this set of three faces. By convention, the corresponding codon is mapped inside the triangle on the first dodecahedral face, and in the reading direction of the second face.

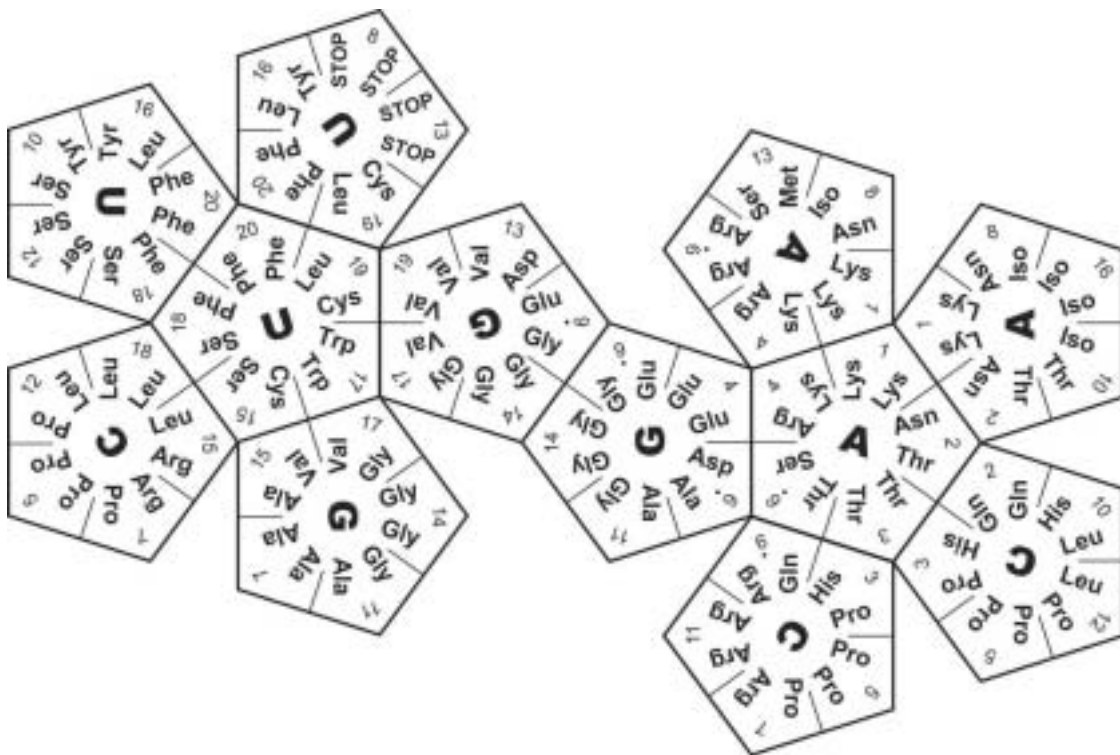


From here we can develop more terminology to help understand the data within this structure. There are three different classes of codons in the genetic code depending on the heterogeneity of the nucleotides in a codon. Each class generates one or more types of codon based on the order of nucleotides. The major poles create four primary triplets that have homogenous nucleotides; UUU, for instance. There are twelve semi-homogenous secondary triplets, such as UUG, and four completely heterogeneous tertiary triplets, such as ACG. By joining mirrored permutations we can enhance the appearance and readability of the map.



Class	Type	Total	Example	Multiplet
Primary	1	4	UUU	Homo
Secondary	1	12	UUG	Homo (3)
Secondary	2	12	UGU	Hetero
Secondary	3	12	GUU	Hetero
Tertiary	1	12	ACG	Hetero
Tertiary	2	12	AGC	Hetero

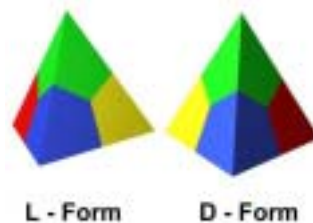
Twelve nucleotides oriented in twenty triangles on a three-dimensional surface create a symbolic network that preserves inherent symmetry in the data of the genetic code.



The network produces a variety of units of nucleotide symmetry that can inform our view of the genetic code. 'Doublet' and 'triplet' are terms to stand for unordered sets of two and three nucleotides respectively. A codon is a specific permutation of a triplet. The term 'multiplet' will mean a group of four codons that all share the same B1B2 nucleotides.

- Four major poles.
- Twelve singlets.
- Sixty doublet permutations.
- Sixteen multiplets.
- Twenty distinct triplets with six possible permutations each.
- Sixty-four distinct triplet permutations representing all codons.

In addition to the reading conventions already described, there is one other significant convention adopted by the Rafiki map at the level of the four major poles. There are only two nucleotide configurations that lead to a comprehensive mapping of codons, and they are stereoisomers, or perversions of each other. They are made interchangeable by swapping any two of the major poles, because the only distinct arrangements of a tetrahedron are mirrors of each other.



Data patterns form in both versions of the map, but given the data provided by nature the one on the left is visibly superior. I will call the superior configuration of nucleotides the L-form, and use it exclusively to stand for the Rafiki map. I use the term as if it is a single map, but note there is a D-form as well. Compared to other structures for the data, this degree of subjectivity is small, and it might even have pedagogic value with respect to fact that all amino acids in the standard set are L-isomers. Beyond this, nucleotides in a dodecahedron must be objectively configured to achieve a comprehensive map of codons.

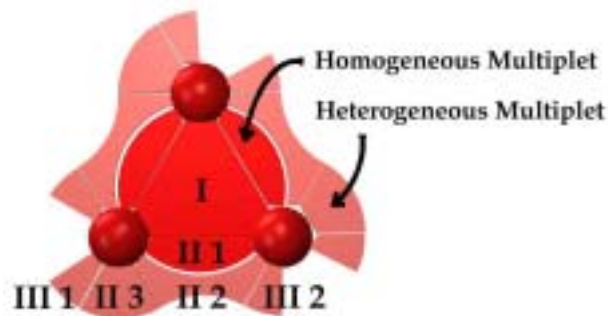
Symmetry in the Rafiki map expands from the poles, creating a hierarchy of elements culminating in logical Hamiltonian circuits of entire nucleotide sequences. The hierarchy of symmetric elements in the structure is as follows.

Pole : Singlet : Doublet : Triplet : Codon : Circuit

Symmetry in the poles has already been addressed within the context of a tetrahedron, and symmetry in nucleotide singlets derives from the faces of a dodecahedron. However, the relationships between individual nucleotide symbols become more informative within the context of the four poles. They offer a convenient method for distinguishing single nucleotides relative to the context of the entire network, because every face on a dodecahedron is parallel to exactly one other face. All four poles of the Rafiki map include three faces, and each of these three faces is parallel to a face from a different pole. We can use subscripts (McNeil subscripts) to denote parallel faces, effectively creating twelve distinct nucleotide symbols.



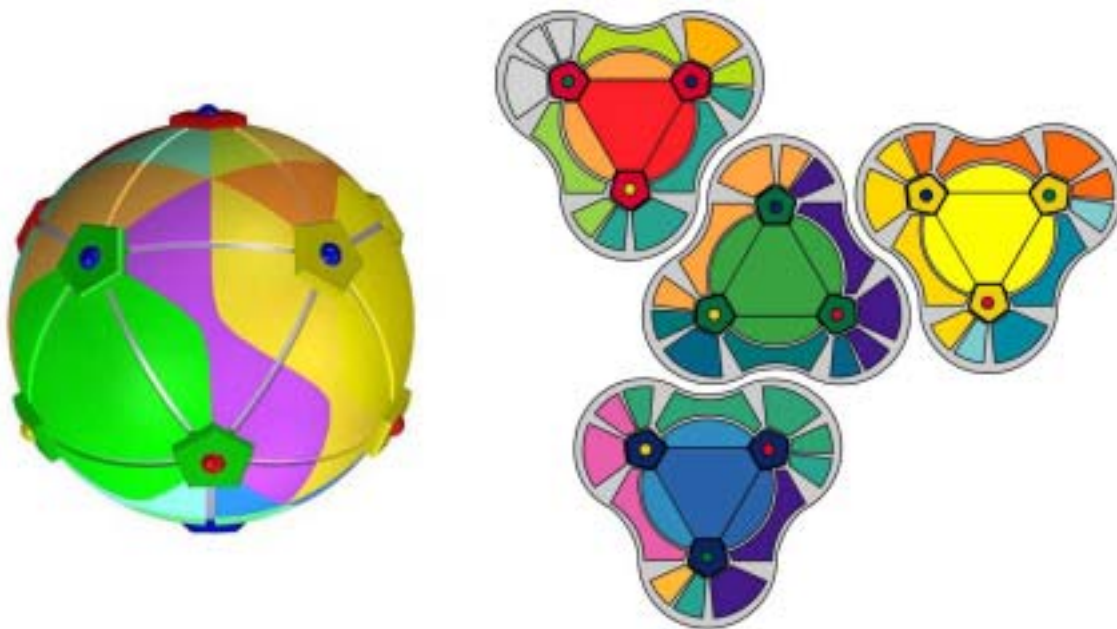
These subscripts help distinguish symmetries in the hierarchy beyond singlets, including the sixty distinct nucleotide doublet permutations in the map. Without unique identification of twelve nucleotides there can be only sixteen distinct doublets. B_1B_2 doublets (base one and base two of a codon) aggregate into sixteen multiplets consisting of four codons each. These multiplets correspond to the B_1B_2 portion of a codon, and they have long been recognized as significant within the assignments of the genetic code. There are two types of multiplets that correlate with homogenous and heterogenous doublets. For instance, the doublet UU generates the homogeneous multiplet of UUU, UUA, UUC and UUG, whereas the UG doublet leads to a heterogeneous multiplet made of UGC, UGG, UGU and UGA. The latter outnumber the former three to one, and together they form a quartet around each major pole.



The quartet of multiplets centered on each major pole is a completely symmetric interweaving of multiplets within the overall structure.



Of course, this data is best viewed in three dimensions, but we usually have only two dimensions available on a printed page. In that case, it is convenient to organize the data along either the faces of a polyhedron, or within the structure of the four major poles. To illustrate, I shall use amino acid water affinities to assign colors within the structure. The red portion of the spectrum colors the most hydrophobic amino acids, and hydrophilic residues are given the blue end of the spectrum. Yellow and green color the middle affinities. Isoleucine and arginine are red-purple and blue-purple respectively to connect the two ends of the color wheel. Colors for the nucleotide faces and their McNeil subscripts are added as well (A=blue, C=green, G=yellow, U=red).



Summary

This short paper is intended merely as an introduction to some of the properties and terminology of the Rafiki Map. It is not meant as a comprehensive interpretation or review of the applications for this data structure. However, the pure symmetry and objectivity of the structure make it ideal for studying patterns in the genetic code. Its advantages include the following.

1. It is the most symmetric treatment of the data.
2. It is the most compressed arrangement of nucleotides.
3. It is the most objective arrangement of nucleotides.
4. It is based on the symmetry of real, three-dimensional objects, and the geometry of these objects is congruent with the geometry of molecules comprising the genetic code.

Due to these qualities, the Rafiki map is useful for viewing and normalizing patterns across the entire data set, or for comparisons of patterns based on a specific triplet, class or codon type. Because the map is based on real polyhedra, as opposed to the imaginary structure of hypercubes and spreadsheets, it can augment studies of various relationships between real molecules of genetic translation. The genetic code is part of a language where a sequence of dodecahedrons (DNA) communicates information to a sequence of tetrahedrons (proteins). It is no accident then that this language fits so perfectly into the structure of a dodecahedron. The map's perfect symmetry is useful in visualizing the effect that codon assignment patterns have on translation, especially in the case of frameshifts.

Beyond all that, the harmony of spatial relationships within the structure, and the aesthetic appeal of its simplicity are likely to prove informative to our views on the origin, logic and evolution of the genetic code. More importantly, this structure might shed light on functions of information handling that take place during genetic translation. This is a process with which we are still remarkably unfamiliar, despite all of the past hyperbole. I must confess that the data when studied in this structure has completely infected me with the heretical conviction that information translated by the genetic code is primarily about protein structure, not sequence. After all, the code is directly responsible for the peptide bonds that make up a native protein, so it is logical to expect the code to be optimized for that function. The genetic code builds proteins, not sequences. Fear not - that argument will not be made here.

In summary, the Rafiki map is a useful general tool for studying the data set known as the genetic code, because it brings maximum symmetry to its structure. The beautiful symmetry in the genetic code need no longer be imagined; it can now be held in the hands and seen with the eyes.