

Why Three?

Making Simple Numerical Sense of Complex Molecular Sets

By Mark White, MD
Copyright 2007, Rafiki Inc.

“Truth springs from argument amongst friends.”

David Hume

Abstract

It is accepted that the nature of DNA determines the first principles of its translation. In other words, DNA is a linear sequence of only four different kinds of nucleotides. This appears to mean that DNA must be organized into sets of three consecutive nucleotides – codons - and then related to twenty different amino acids during operations of the genetic code. However, this simple logic is no longer as obvious as perhaps it might seem. In fact, this logic is probably inverted. It seems more logical now to think that DNA has been organized at bottom by codons, as well as the broader structure and functions of the genetic code itself.

Codons First

The genetic code, as it is defined today, is embodied in a standard codon table. This simple spreadsheet, or a data look-up table, shows us the assignment of twenty amino acids to sixty-four codons. Since there are but a handful of known exceptions to it, the standard codon table is - for all intents and purposes - universal to all life on earth. A codon is a set of three consecutive nucleotides found in a messenger RNA molecule (mRNA). There are four different kinds of nucleotides in mRNA, like in DNA, and for the most part, codons can be thought of as being “stored” in DNA, only to be transcribed into mRNA. Most amino acids are assigned in the table to more than one codon. In reality, amino acids are only associated with codons via transfer RNA molecules (tRNA). There exists a sequence of three nucleotides in the tRNA - called an anticodon - that is responsible for recognizing codons.

Amino acids form sequences via peptide bonds. Nucleotides form sequences via phosphodiester bonds. Codons form a sequence in DNA. Amino acids form a sequence in protein. We now say that the genetic code is the “linear” relationship between sequences of DNA and protein, and the entire logic of it can be found in the codon table. However, there is no logical, organizing structure to the table itself, and no bond information is contained therein. It is strictly an arbitrary arrangement of linear data relating nucleotide triplets and amino acids. The genetic code today merely charts the near-universal assignments of amino acids to codons. A codon today “means” an amino acid in the genetic code. It is strictly a numbers game.

George Gamow is credited with first determining that codons would be made from three nucleotides¹. He quickly determined this because he already knew that there were four nucleotides in DNA and there are twenty amino acids in the code of their translation. From there it is pretty simple math, really, because $4 \times 4 = 16$ and $4 \times 4 \times 4 = 64$. There must be at least 20 codons – perhaps - but was his basic numerical argument correct? I think not. I can make this outrageous claim now because Gamow’s simple argument was clearly based on missing information and at least one critical premise that turns out to be false. Consider that at the time Gamow knew nothing of mRNA, tRNA, peptide bonds and protein complexity. Does it matter? I think it clearly does. The three nucleotides in a codon must have logically preceded the four kinds of nucleotides in DNA, as well as the twenty amino acids in protein. After all, it is still just a numbers game.

The genetic code involves sets of molecules and the logical relationships between those sets. The sets of molecules are remarkably consistent, yet the molecules themselves have no way of “knowing” that they are actually in sets. Molecular environments are generally chaotic, yet the insentient molecules participating in the genetic code are notably well-behaved across many scales of time, space and various life forms. This strongly implies to us that there are

optimums involved in the genetic code. There are optimum molecules, optimum sets and optimum relationships between these sets. What are they? What are the molecules, what are the sets, what are the relationships, and what are the optimums? It is obviously still a numbers game.

The two basic sets of molecules in life are nucleotides and amino acids; however, these molecules must be bonded together to form more sets, larger sets, and more complex sets of more complex molecules. Molecular bonds are therefore important. They are an essential part of translation because without bonds there is no genetic code. The nucleotides are bonded together by phosphodiester bonds that tend to be fairly homogeneous. Their stereochemistry and their rate of formation appear to be mostly irrelevant to the molecular information in the more complex sets that they make. Amino acids, on the other hand, are bonded together by peptide bonds, and these bonds do seem to be informative to the system in generalⁱⁱ, ⁱⁱⁱ. Their rate of formation and their stereochemistry *are* relevant to the relationships between complex molecular sets in both time and space. From these simple observations we can make a broad list of the sets of molecules that might reasonably be seen to participate in the operations of the genetic code.

Molecular Sets Broadly Involved in the Genetic Code

- Nucleotides
- Amino acids
- DNA codons
- DNA sequences
- mRNA codons
- mRNA introns
- mRNA exons
- mRNA sequences
- Anticodons
- tRNA
- Aminoacyl-tRNA synthetase (ARS)
- Amino acid sequences
- Peptide bond sequences
- Proteins
- Protein aggregations
- Whole protein populations

When we talk of molecular information we must think in terms of sets. The information of each molecule ultimately will be derived from the notion of “all possible” molecules in the set to which each molecule belongs. When we talk of

codes, or molecular languages, we must think in terms of logical relationships between sets in time and space.

The genetic code exists only as a function of many sets of molecules in time and space. All molecules are independent of other molecules in their own general actions because all insentient molecules are ignorant of the actions of other molecules. However, the molecular codes are always dependent on collections of molecules and their collective actions. Collective action defines molecular codes. We can say that any reliably repeating specific collection of molecular sets and molecular actions in time and space must constitute a molecular code. Therefore, the genetic code does not exist independent of any collection of molecular sets. It is purely a function of a specific set of molecules in space through time. The information content and the codes that lead to a formation of molecular sets is a function of the specific molecules, their numbers, and their specific logical relationships. Different sets of molecules will necessarily have different codes. Every instance of the genetic code today is a collection of molecular sets that results from evolution of molecules and molecular sets.

All of the subtle varieties of the genetic code today are the result of evolution. Complex molecular sets have evolved through time on earth. The molecules, their sets, and the logical relationships did not exist at the origin of earth. They all have evolved in concert since then. Some of the sets in the genetic code are nearly universal on earth, but this does not necessarily mean that all of the codes are exactly the same - far from it. Some of the molecular sets involved in the many different genetic codes vary considerably between organisms. However, all of the sets and codes - being products of evolution - are tending toward optimums. Our task then is to identify these sets, recognize their information and logical relationships, and then try to infer their optimums. This is roughly what Gamow did, but he was woefully ignorant of most of the relevant sets, the relationships between them, and so he based his simple conclusion on demonstrably false premises. His initial premise was that there is a starting point of four nucleotides and this now is *demonstrably false*.

The set of codons is first logically related to the set of anticodons during translation. Codons "mean" anticodons when proteins are made; they do not "mean" amino acids in the time-dependant natural function of the genetic code. Granted, there are virtually always three nucleotides in all codons and all anticodons, but there are also more than four kinds of nucleotides in the set of anticodons. This means that there are logically more anticodons than there are codons. This in turn implies that the set of four nucleotides in DNA is derived from and optimized for the fact that there are three nucleotides in a codon, and not vice versa. The genetic code must first make a logical relationship between DNA and tRNA, and this is now dependent on there being three nucleotides in a codon, not four kinds of nucleotides in DNA. Our immediate concern should now be to determine the basis and origin of these numeric relationships. After all, the grand function of codons is to relate sequences of DNA to proteins. This must be

taken in the context of all possible DNA and all possible protein. The optimums in these much larger sets apparently evolved as a consequence of the logical relationships derived from a premise of three nucleotides in a codon and not from a premise of four kinds of nucleotides in a codon or anticodon.

Codons are translated into anticodons, which are translated into sequences of tRNA. Sequences of tRNA are directly translated into sequences of peptide bonds. This is logically true. Any statement that there is only a direct translation from tRNA into sequences of amino acids is spurious because it ignores the absolute necessity of making a specific peptide bond between every two amino acids. Every sequence of amino acids must first be formed by a specific sequence of peptide bonds that must result from translation in time and space via consecutive tRNA pairs. The translated sequences of peptide bonds then determine the set of all possible proteins within the system. The set of all possible proteins, in conjunction with DNA, then determine the set of all possible sequences of DNA. In other words, the same molecular information that is translated from DNA into protein must be used to make more DNA.

By making optimized molecular sets, the entire language becomes efficient and the sets become extremely biased. Understanding this process, I think, begins with recognizing that there are three nucleotides in a codon. This is where it all starts. This optimum number then determines the numbers - moving downward - for DNA, mRNA, and - moving upward - for tRNA, peptide bonds, and ultimately it must determine the numbers for proteins. These sets, although much larger and more variable, must also represent optimum sets. Four nucleotides is apparently an optimum for storing codons, and twenty amino acids is an optimum for translating all codons into the most efficient set of all possible peptide bonds. Sometimes more bonds are required than others, so sets of tRNA vary greatly. The amino acids stay the same, however, and so too do the sets of ARS.

The correct starting number now apparently is three nucleotides in a codon. From this, an optimum number of nucleotides in DNA is four. This is the optimum for efficient storage of symmetrical sequence and structural information within this particular system of molecular information. Twenty amino acids capture this symmetry for peptide bonds, which are derived from those biases created in nucleotide sets. Three nucleotides in a codon provide the highest possible symmetry group and the most effective system of transformations within the entire codon set. It is both effective and efficient. This is essential toward efficiently making symmetrical sets above and below codons in the hierarchy of molecular sets. Some organisms prefer to make more distinct kinds of peptide bonds than do others, so there is a large degree of flexibility in sets of tRNA between organisms. This flexibility is not seen in sets of amino acids and ARS, but it can be seen in tRNA and peptide bonds.

If we want to begin to make simple numerical sense of the complex molecular sets operating the genetic code in nature we must first realize that codons provide the primary logical basis for that organizational structure. Codons organize DNA, not vice-versa, and codons organize the sets above it, starting with anticodons, tRNA and peptide bonds. After all, it is just a numbers game.

Contact: Mark White, MD mark@codefun.com

References

ⁱ **Gamow, G.**, "Possible relation between deoxyribonucleic acid and protein structure.," *Nature*, vol. 173, pp. 318, 1954.

ⁱⁱ **Cortazzo P, Cerveñansky C, Marín M, Reiss C, Ehrlich R, Deana A.** Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochemical and Biophysical Research Communications*. Volume 293, Issue 1, 26 April 2002, Pages 537-541

ⁱⁱⁱ **Kimchi-Sarfaty C, Oh JM, Kim I, Sauna Z, Calcagno AM, Ambudkar S, Gottesman M.** A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*. 2007.