On the Proper Approach to Modeling Molecular Information and the Genetic Code

By Mark White, MD
Copyright Rafiki, Inc. 2006

“It takes a long time to become an overnight success”

Jeff Parker

Introduction

Science today lacks a workable definition for molecular information. Finding a proper definition is perhaps the single greatest scientific challenge of our time. Within all of biology, our current definition of molecular information now only addresses the “linear” arrangement of sub-units of any larger bio-polymer. The information of any molecule is found in its sequence, just like this sentence is a sequence of letters or just like a string of computer code is a sequence of zeros and ones. This “one-dimensional” definition wrongly excludes all other forms of information. Unfortunately, the true nature of informative molecules is currently unknown, but we should know - without any doubt whatsoever - that it is in reality a multi-dimensional entity. In other words, our current definition of molecular information is a logical subset of real molecular information. What then is the proper superset? Any useful answer will require an entirely new approach to the problem.

The intensity of the current problem is magnified by knowing that many other basic concepts in molecular biology are in part derived from or related to this one fundamental definition. We talk freely about genuinely abstract notions of “genes,” “translations” and “codes,” all in the general context of molecular information. But what happens if this context turns out to be utterly false? In other words, what if we have simply failed to grasp the first principles or the very nature of the genetic beast itself? Our failed model of the genetic code is a good example here of the looming problem in the future advances of life sciences. The genetic code is only that code that translates molecules one way; from DNA into proteins. It should now be seen as an absurd, backward ghost of reality, simply the unfortunate derivation of a flawed definition of molecular information in the first place.
Consider that even the most simplistic view should respect the fact that -
on the smallest scale - molecular information must be a real, quantifiable entity
that completely defines a particular molecule. It must embody an actual, specific
molecular configuration and make it entirely distinguishable from a set of all
possible configurations. Think of it in terms of names - every distinct molecule
must not only have a name but it must somehow also be namable. Like
electrons in an atom, where all electrons are considered identical yet no two can
be the same, so too is the conflict in our perception of molecular information
created by not having a simple means to make sets, or to make sets of
distinctions. The simplest possible concept of molecular information alone will
never meet our conceptual needs here because, on any larger scale, molecular
information is a tremendously complex and highly recursive entity. It defies any
simplistic human understanding. The genetic code represents a natural
language of molecules. We must expect then that it will somehow rival the
complexity of all human languages combined.

There are, in fact, many potential facets to any form of molecular
information, and it is therefore quantifiable only through a comprehensive
understanding of complex molecular relationships. The whole issue of identifying
molecules by their inherent “information” must be taken within a broad context of
any specific environment, yet also within the context of all highly variable
molecular environments. It must also always be put into a relative scale of time.
Molecules are merely manifestations of thermodynamic events, a term which
literally means heat motion. Any kind of motion is purely a function of time and
space. Therefore, it is folly to think that molecules can be defined by a simplistic
entity that excludes both time and space from its very own definition. Every
counting of “possible molecular configurations” for any particular molecule is
entirely dependent on the specific context for that molecule in time and space.
So, man can review his work to date, and we can conclude that we have
heretofore made the simple mistake of thinking of molecules as we think; when
instead it is rightly a question of how they think. Therefore, to advance our
understanding of molecular information we must better our understanding of how
any molecule must think.

Let’s start with a simple definition of information; it is the counting of
choices. Counting can only be done in whole numbers of real choices, so there
must first be things to count – choices to make – and then there must be ways of
making choices. Ultimately, any form of information is quantified by also knowing
the probability of choices made. Plus, probabilities are always related to the
specific mechanisms for making those choices. What are the specific choices
and the mechanisms for making definitive choices in the insentient world of
molecules? When we can answer that difficult question, we will have our elusive
definition of molecular information.

What is a choice? We can answer that most simply in terms of pure logic
if we recognize that counting produces sets and sets can be logically related. A
logical relationship between sets is called a function. In these terms a choice is
the operation of a function. In other words, a choice is always made based on the logical relationship between things in sets. A function in these purely logical terms is always symmetrical to some degree, or it somehow works both ways between sets. The more perfectly symmetrical a function is, the more completely can information be translated back and forth between sets. All languages are codes and all codes are functions. When we enumerate the steps in the operation of a function we have produced an algorithm, which is the specific list of steps of a function. When we write, read and execute algorithms we are working with codes. There are many ways to write, read and execute any code. A language is a specific way to execute a code. In other words, for any function there can be many possible languages. We can generally say then that choices are logically made through the execution of languages. What is the language of the genetic code? When we can answer that, we will know the first principles of life itself.

Life is a process that somehow organizes molecular information through time. The language of life must, therefore, be a language of organization. This simple idea merely reflects a basic definition of evolution itself, where evolution of any kind is the abstract notion of organizing information through time. Life is simply a manifestation of evolution in general, and life represents an organization of molecular information through time at larger and larger scales, involving evermore-vast periods of time and volumes of space. However, any specific organization of molecular information is itself an important form of molecular information. It stands to reason then that since languages are forms of organization, molecular languages now must also be seen as complex forms of molecular information. The genetic code has rightly come to be our central paradigm of molecular information, but our overly-simplistic model of the genetic code today is inadequate toward our understanding of life’s many patterns of organization, and this is true for several obvious reasons:

- It is based on an inadequate definition of molecular information.
- It denies that sequence is always a logical subset of structure. In fact, it inverts this logic along with all explanations of it.
- It fails to recognize the existence of any primary, logical organizing structure for molecular functions in general.
- It fails to appreciate the required roles played by time, symmetry and molecular context that must be integrated into the function of the genetic code at all levels.
- Neither the basic sets nor the basic functions have yet been identified in any workable definition of the genetic code.
- The genetic code is purely molecular, and it is at first a structural function built of structures used for its inputs and outputs, as well as the fundamental logic of its operation. Therefore, no model that eliminates every trace of structure from the genetic code will ever adequately explain
the genetic code. Molecular information is always at first structural information, and so its translation is always a structural function.

- The genetic code has tightly organized molecular information on earth through time, obviously, but the standard model comically sees this primarily as a “frozen” and “arbitrary” event devoid of and immune from evolution rather than a significant hallmark of a broader, logical (i.e. not arbitrary) process of evolution.
- The genetic code is complex not simple; it involves many facets of molecular information not just one, and it demonstrates a coherent organizational structure that works broadly to expand molecular information through time not compress it. Therefore, it must be understood within a broader context of the entire molecular system of life through time.

For these and other practical reasons, our basic model of the genetic code must change along with a search for a new, workable definition for molecular information. Our current model simply does not work for any of the complex intellectual tasks we now need it to perform.

**Idealizing the genetic code**

In 1953, Watson and Crick published their famous letter to Nature proposing a double helical structure for the DNA molecule. This was the shot heard round the world of molecular biology. The simple structure was taken as an idealization of the DNA molecule, yet it is widely acknowledged that the molecule itself can exhibit many subtle and not so subtle variations of this structure. In fact, there have been many credible and useful challenges since then to their basic proposal, as well as challenges to the logistical implications of this. However, the idealization of this structure has rightly withstood most if not all of its ideological challenges. It has proven to be a “good” idealization. It still serves as the starting point for our understanding of molecular information, as well as the basis of many simple human symbols and languages that we use in trying to describe and understand the natural role of DNA in living systems.

At the time, there existed no accepted formal idealization of the genetic code. It is natural then that DNA would serve as our starting point for idealizing the genetic code. However, the stunningly simple and seemingly consistent structure of DNA - when seen as the primary storage mechanism of molecular information - also strongly implies that its translation would somehow involve a simple mechanism, one that could perhaps be completely devoid of any significant structural elements. The specifics of DNA and its idealized structure were then logically used as the foundation of our initial attempts to idealize the mechanisms of its translation. In retrospect, this was a big mistake.
Unfortunately, all molecular codes are bound to be more complex than any of the individual molecules that participate in their execution, and they are logically bound to involve complex relationships between all of the molecules that comprise the codes themselves. After all, molecular codes are merely logical relationships between molecular sets through time. Note that all molecular information is at bottom structural; even “simple” molecular sequences are sequences of simple molecular structure. For molecular information, structure always logically subsumes sequence. This basic and logically correct concept was completely lost in the initial efforts to idealize the genetic code; therefore, the natural function of the genetic code is now entirely inconsistent with our current idealization of it. This is not a good situation for science to find itself in today, and so there is much foolishness now on display. We have idealized this important code as a fundamentally unstructured entity that merely translates sequence. Time and structure are eliminated from the model entirely. The function of the genetic code is indisputably now only viewed – entirely defined actually - as the relationship between a set of codons and a set of amino acids. The relationship is embodied in the codon table alone. The icon and its relationship to time and structure are still further described by the central dogma of molecular biology. We can now immediately understand all of the relationships between all of the elements in all of the molecular sets of the genetic code via the following simple icon of a flow diagram:

Figure 1.

This graphical icon completely defines our current idealization of the genetic code. There can be no real dispute about that. It illustrates the location and function of the genetic code, an entity devoid of structure and structural function. The icon shows the code’s relationship to time and structure; it is one completely separate from either of those things, and free of all functions that operate upon them. While it is true that an important function of the genetic code is to relate codons to amino acids, it is also true that this is not its only function in nature. Furthermore, our goal should not be to idealize things in the simplest possible way, but to idealize them in the simplest possible way that is also the most correct, and in a way that promotes the least confusion. For instance, we do not idealize the earth as a circle even though this is the simplest possible way to do it - simply because the earth is known to have volume. We can idealize it as a sphere, to some degree, but even this is not entirely correct because the earth is known to be slightly irregular and ovoid. Still, moving from circle to sphere is a tremendous improvement. Likewise, we cannot idealize the genetic code as a codon table because we know that many important sets and elements
are missing from this model. Moving from sequence to structure is a tremendous improvement in our model. Therefore, we must definitively move away from this ancient flat model and start the process of idealizing this natural code in a more robust and rational manner. This will be extremely difficult - much has been invested and many have been indoctrinated - but we can start by recognizing the things we clearly know to be true about the genetic code in general.

- The genetic code is not universal.
- The genetic code cannot be understood as a single, simple, unidirectional function but as a complex network of symmetrical functions.
- The genetic code is a function of space and time.
- The genetic code logically relates sets of molecular structures through time.

The genetic code on earth has evolved. Like all things on earth - especially languages - the genetic code has evolved many variations. In fact, the genetic code involves many functions, and each of these can be seen as having many different languages. So, naturally, there are many combined dialects of the genetic code operating on earth today. Only a tiny portion of these variations is reflected in the known variations of the codon table because the codon table reflects only a tiny portion of any version of the genetic code. So, the task of defining the genetic code becomes more difficult; first because it involves defining molecular information, and second because it involves defining the specific sets that participate in the various functions across multiple languages.

Despite these obvious difficulties, all variations of the genetic code can share the same primary logic, and this accounts in part for the extreme consistency in basic sets and operations seen in all of the languages and their many current variations. The primary logic of the genetic code involves the logical relationship between all spatial structures in the universe. All basic structures in the universe share the same logical relationships with all other structures through shared common geometry. This is the smallest possible scale for our thinking on the problems faced by molecules; but on the largest possible scale, the function of the genetic code is to determine the logical relationship between all possible proteins and all possible nucleotides through time. This relationship must be symmetrical, because proteins and nucleotides are mutually dependent, but this does not mean that the language is the same in relating the two sets to each other through time. In fact, the languages must be quite different in both directions because they are time-dependent functions. Although their information can be the same on a primary level, their information is defined by specifics in both molecular sets, and those sets are quite different in their specifics. However, the set of all possible proteins is at first a set of molecular structures; likewise, so is the set of all possible nucleotides. The molecular information in each set is logically related to the other, but there are an infinite number of languages that are capable of performing these translation functions.
Life is about finding optimums, and life has apparently been quite busy finding optimums in these functions.

Life appears to find optimums best by strongly biasing sets. Note that the set of all possible proteins in life today is an infinitesimally small subset of all possible proteins ever. The same is true of nucleotides. So, life is performing its functions by restricting itself to small sets that are optimized for generating larger sets. The set of all possible codons has been greatly reduced by life’s selection of mostly only four nucleotides. All possible proteins are similarly reduced by life’s selection of twenty common amino acids. Still, the number of structures that can be generated through logical relationships between these sets remains staggeringly large. Life chooses relationships based on efficiency. It uses the fewest parts to generate the most wholes. Symmetrical relationships are the most efficient relationships, especially when it comes to generating molecular structures in time.

A Language of Structures

Before we can begin describing any natural language of structures, we ourselves must have a language of structure to describe it. Please consider that our concept of the genetic code is merely a metaphor. A metaphor is an implied meaning of one thing to another given a particular context. I believe that no example from the English language is necessary here because, according to Dr. Roy Wagner, all human languages are examples of recursive metaphor. In modeling the genetic code we use human language as a means of comparison to a molecular language. Now consider that the genetic code itself is merely a metaphor. It is a molecular metaphor that uses molecules as a means of comparison to other molecules. Molecules must have a basis for comparison, and humans must now have a basis for describing that comparison.

Any human language of spatial structure is called geometry, and there are an infinite number of possible geometries; therefore, we must choose wisely. Our goal here is to describe molecular codes in terms of efficient, symmetrical functions of time and space. Any description of time and space is a description of a universe. So our geometry must have easy ways for us to count elements of time and space in a universe of molecular information. I will describe a simple geometry here that I feel is most adequate for that task. Of course, it reflects my prejudices toward a particular set of beliefs about that universe, so here is a simple list of those beliefs about time and space as it relates to molecular information.

- They define each other.
- They are maximally symmetrical.
- They are purely informative and therefore inherently binary.
- They are self-organizing entities.
Without a protracted philosophical debate, I believe that we can use these ideas to quickly build the most efficient geometry for describing the fundamental structure and logic of the genetic code, or at least in quickly building a structure for a language to describe it. This is an admittedly quirky geometry invented solely for communicating a “solid-state” view of a molecular information universe. Space is at first defined by time, and then space is entirely defined by a collection of combined elements, such as planes, angles, lines and points. The relationships between these elements are purely logical, and those relationships are combined to form more complex larger elements of structure. We will need symbols for the most basic elements, and then all other elements can somehow be made of logical combinations of those symbols.

All geometries are imaginary. In other words, the primary elements for building any geometry must at first be imagined. Therefore, all of man’s languages of time and space are merely used to describe imaginary universes. This is the nature of man’s relationship to time and space. But it has proven to be decidedly useful path toward understanding and predicting the behavior and phenomena we can empirically witness in our many universes.

This particular geometry starts with the abstract notion of imagining time. Time is a purely imaginary concept here, to be sure, but it must be a binary element in this particular informative universe, so we will assign it the symbols positive (+) and negative (-). Time is infinite and time is perfectly symmetrical in this imaginary form. We can now imagine the infinity of positive related to the infinity of negative, and what we will see in our mind’s eye, hopefully, is a perfect plane between positive and negative. So, the plane - not the point - is the primary generative element of this particular geometry. Time is defined as change, and in this case it is simply any change between positive and negative - so too are the initial notions of distance and direction. The symbol one (1) will stand for minimum change and the symbol zero (0) will stand for no change. We can define a plane as the symmetrical interface between a minimum change between positive and negative. Unit time, distance and direction are a function of minimal transformations in either direction between a positive and a negative element in any plane of time. We can now stack planes of minimum change infinitely in the positive and negative direction and thereby begin counting time and distance as functions of an imaginary time and space.

To make this geometry useful for counting all elements of space, we must first create them. We will do this merely by adding new planes to create logical combinations of planes forming angles, points and volumes. Here, an angle is the relationship between two planes. A line is the intersection between two planes. A point is the intersection between three planes. A line segment is the intersection between four planes. An angle segment is the intersection between two line segments. Note that all angles are the same but all angle segments are not. All lines are “straight” but a combination of line segments perhaps is not. A volume is the space defined by a minimum of four points. To fully create this
system of geometry we now merely need to pick a standard unit of common angle between planes. The standard choice in most of our standard geometries is to choose the angle between any two planes in a cube. This standard “Cartesian” choice is a convenient human choice because it makes all angles 90 degrees, and all compound angles are multiples thereof, so angles become maximally easy to understand, and planes become completely filled by similar squares. However, if we choose this as our standard angle then we must know that no perfect dodecahedron can ever exist in this universe, and we certainly will need at least one of them eventually. As Plato said, the dodecahedron is the cosmos and it embroiders the heavens.

This eccentric geometry is, as I previously implied, a form of “quantum geometry” in that it is purely informative. To be purely informative a universe can only have whole units of every element, and only whole unit multiples for all composites of elements. In a purely informative universe, anything that does not exist in whole units does not exist at all. If we hope to ultimately use this geometry to define an information system it would be nice to imagine that we must only have whole choices of things to count, which can then be logically arranged into sets. We should thereby choose sets that lend themselves to making the most efficient combinations of sets in this regard. Ultimately, we hope not just to make simple sets, but to make complex sets of sets of sets… up to the largest set in the hierarchy, which is the whole universe.

If we choose cubic symmetry we must sacrifice the dodecahedron, but if we choose the dodecahedral angle for intersecting planes we will generate more than enough cubes. In other words, if we were computers we would choose the dodecahedron every time. Since we are men, we consistently choose a cube. Men are like that. We now must ask the question: What would a molecule choose? After all, we are building this particular geometry to communicate ideas about a natural molecular language. I think a molecule, like a computer, would choose the dodecahedron for use in a system of maximal symmetry, efficiency and power with respect to spatial information and its logical functions. Therefore, this quantum geometry will be built on the six symmetrical planes of a dodecahedron. You can label them A, B, C, D, E and F if you want, or you can color them red, yellow, blue, orange, green and purple. You can assign them each a note from a perfectly balanced musical scale. It does not really matter what symbols you use because these six planes are all the same relative to each other. They combine infinitely in all directions to produce the temporal and spatial elements of the universe that we must have in some form before we can start counting elements and sets as whole numbers, before we can start producing any logical relationships between elements and sets.

All of the spatial structures can now be built via translations in time within the context of this quantized geometry. We can start with the quantized unit of space, which is now a dodecahedron. It is formed by the intersection of all six primary planes. There is a positive and a negative surface to every plane, twelve distinct surfaces are created, twenty points and thirty line segments. All of the
combined elements of this geometry are maximally efficient for making elements and sets that are binary, quantized and perfectly symmetrical. This seems like a heck of a good way for a molecular system to found a language used in building and searching spatial structures through time. The system can become evermore complex through time, but the primary logic can remain the same. This logic merely expands into more structures and more complexity between combined larger sets of parts of these same basic structures.

Note that our initially abstract notion of binary directionality of time in space becomes distorted by the addition of five symmetrical planes to our original plane of time. Time no longer has a single clear meaning, yet it remains relative to and symmetrical with space, per Einstein’s instructions. Unfortunately, it becomes inherently directionless in space in a purely quantized way. We might even consider the notion of adding a dual set of planes to the original dodecahedron to make the entire structure perfectly symmetrical with respect to time. Regardless, a global arrow of time remains within the more general context of scale. Time becomes a pure metric of scale with respect to quantities of change in space through time. There is a constant between change in space and change in time for all structures in this universe, and scale is a function of this constant.

Since this universe is self-organizing, the arrow of time now must point toward an expansion of structural information in quantity and complexity through scales of time. The actual universal metric for scale in this universe can now happily be located and identified as phi, the mystical golden ratio, which here becomes the scaling length of a unit line segment relative to the unit distance. Spatial structures are now logically related to time quantities by whole multiples of phi. Large structures are logically related through larger quantities of time than are smaller structures. Larger units of time produce more spatial information through the inevitable self-organization of structures in space in this informative universe. Time can now be measured in terms of information content and complexity of spatial information. This avoids conflicts with basic principles of relativity but now puts us in obvious direct conflict with a somewhat vague notion of entropy. Specifically, this points our time arrow in the opposite direction to those involved with understanding the cherished notion of the second law of thermodynamics. This physical “law” tells us that entropy should be increasing and not decreasing through time. In other words, structural information should perhaps be decreasing universally through time. However, this seeming conflict can be easily rectified if we imagine that this particular universe starts in a state that is moving at maximum velocity toward structural entropy. However, all measured changes in entropy in this universe are a function of a self-organizing force producing acceleration, or in this case deceleration of structural entropy. In other words, the second law should be seen here as one that describes a curve rather than a line relating time and entropy. An increase in entropy and a change in increase in entropy are logically related entities, but their arrows can point in opposite directions.
The primary structural logic of this quirky space is based on fundamental sets of structures. The primary structure in this space is a dodecahedron, but within this one structure there are many logically related substructures, like cubes and tetrahedrons. There are five full cubes within the set of points of a single dodecahedron, each cube sharing two points with the other four cubes. Each cube is composed of two tetrahedrons sharing no points. So, each dodecahedron is composed of ten tetrahedrons all logically related to each other; consequently, any set of tetrahedrons is logically related to any set of dodecahedrons, and any set of dodecahedrons is logically related to any set of tetrahedrons. The dodecahedron, in other words, is a natural and entirely logical metaphor for cubes and tetrahedrons. There are countless possible logical functions for translating the specific elements of complex tetrahedrons into the elements of complex dodecahedrons - and vice versa. This is merely a result of the fact that tetrahedrons are informative compressions of dodecahedral elements. It is not “as if” a tetrahedron is a dodecahedron; a tetrahedron really is a dodecahedron, only with certain parts missing. So, there are an infinite number of possible languages that can exist for performing these translational functions. Life will surely start by using the logic of these naturally existing functions and then quickly find an efficient set of molecular languages for performing them.

We can ask about the structure of some specific molecular phenomenon, like table salt, for instance, and we can find a structure in the universe that demonstrates it. In the case of sodium chloride, a cube demonstrates its fundamental molecular structure. We can ask about the structure of another crystal, like diamond, and in this case it is the tetrahedron that demonstrates its molecular structure. We can ask about how these two structures are related to one another, and this question can always be answered by simple mathematics to show these spatial relationships as well. However, what shall we do when we ask about the structure that relates all structures to all other structures? Does such a thing even exist? Apparently it does, and apparently life has been demonstrating it to us over the past several billions of years. Life is the ultimate metaphor of molecular structural metaphors. It is the living language of crystals.

We can get a quick and dirty understanding of the nitty-gritty mechanics of such a process by briefly returning to our Cartesian roots. Imagine that the universe is a Rubik’s Cube. Imagine that the universe starts out in a maximally random state, and that time is a simple sequence of quarter turns that organizes cubic information in three cubic planes. A self-organizing universe will eventually “solve” the cube via a sequence of logical moves that relate one state to another. All states are logically related and all sequences are logically related. Time becomes a function of the number of turns past present and future, as well as a measure of the “order” produced by any sequence actually produced. Of course, the complexity of this analogy is low in large part because there is only one possible structure in this Rubically informative universe of minimal structures and simple time sequences. This particular cube can only make other informative cubes, whatever they may be. However, the universe of molecular information
operates on the complex relationships between many different structures and the logical relationships between them through time, on scales both small and infinite. Furthermore, persistent and identifiable patterns inevitably must develop on all scales. It is an admittedly coarse analogy, but we must start with something we know before we can begin to understand the things that we do not know. That is the metaphorical nature of human thought.

Sequences of dodecahedrons can always be logically related in time and space with sequences of tetrahedrons. Building, searching and moving structures through time and space is an inherently logical process, even if it is somehow driven by “random events.” However, this process of spatial organization can obviously be simplified by using consistent elements of highly symmetrical, pure structures. This is precisely what life is doing. Life seems to be organizing many sets of many elements of many dodecahedrons through time. The elements and sets themselves are perhaps becoming evermore simpler, but combinations and relationships between them are growing evermore complex. Consequently, structural information is accumulating through time. Complexity is increasing. New languages are spontaneously forming to translate and amplify the new information and complexity that is inevitably forming in this molecular information universe. We can graph and perhaps measure the time course of complexity in this universe by merely examining the time course of life on earth. It is increasing and accelerating. If we are going to logically understand this we must start with logical models and languages that help us do this, and choose only those that add the least possible confusion.

This is admittedly a toy model of cosmology based on a decidedly idiosyncratic flavor of geometry and physics. It is perhaps not useful in any way for understanding basic math or physics, but it now appears to be indispensable when it comes to first understanding the molecular sets in life. To start we must understand them in their simplest possible terms, in terms of information, time, structure and the inherent logic of interdependent molecular sets. We can immediately get started here by using this quantum geometry to neatly rearrange the molecular sets represented by our old friend, the codon table. Nucleotides can be idealized as the faces of a single dodecahedron and amino acids as a symmetrical collection of tetrahedrons. By symmetrically placing three instances of only four kinds of nucleotides into a dodecahedron, we have broken the perfect symmetry of that structure, but we also have generated a huge number of new symmetrical sets. More pragmatic is the realization that we efficiently generate all required sets of common codons into which we can find a home for all common amino acids. They are all now examples of elements in sets of more sets of the exact same thing: they are all parts of a dodecahedron. The codon table in this new light does not define or even describe the genetic code, but it does illustrate the primary logic of its general organizational structure. The genetic code does not operate only on these simple molecular sets but on a more complex logic between much larger and more complex molecular sets through time and space. This primary logic merely reflects the most efficient way
to organize sequences of dodecahedrons into more complex and spatially informative sets of the elements of larger dodecahedrons. Many natural languages of structure are logically formed within this particular system, and they compete endlessly to become the basis of new, more complex languages.

People become understandably confused here. They seem to want to find this molecular structure somewhere in our universe. Be ever mindful that no such molecular structure exists now or perhaps ever. I am not necessarily saying that the genetic code is a dodecahedron; I am saying that a dodecahedron is the best possible metaphor for any human understanding of the primary logic of the genetic code. The molecules of the codon table perhaps do not organize into a dodecahedron but the dodecahedron did clearly organize the molecules of the genetic code. Think of it in terms of naming. Before we name we must have structures that require a name. Life self-selected the best possible names for the common elements of a molecular dodecahedron and thereby inherited all of the primary logic and possible languages useful for relating structures to other structures in a multitude of molecular contexts.

We can now happily continue to idealize the double helix as the simplest set of molecular dodecahedral sequences in this particular scale of time and space. We can now also idealize proteins as the simplest set of tetrahedral sequences in this scale of time and space. The genetic code translates these sequences through time, first into sequences of tetrahedrons that must somehow be perceived as hidden symmetrical structural elements of much larger dodecahedrons. These new sets merely combine to form a vastly larger set of complex “dodecahedrons” that ultimately appear as specific proteins or parts of proteins, or more precisely into collections of molecules that have their own consistent macromolecular structures. Yet these structures still must be tasked through time with making more combinations of the spatially simpler and greatly compressed sequences of nucleotides. And thus the symmetrical cycle of logical molecular relationships between time and space continues. Information and complexity expand with each biological tick of the grand universal clock.

Life is an inherently logical process of fractal information expansion. In the prescient words of Erwin Schrödinger, it is producing order from disorder and it is producing more order from order. All of the structures of life that are output from each function immediately act as the input for other functions. Plus, the output from some functions gets immediately input back into those same functions. Choices are constantly being made by the ubiquitous mechanism of natural selection - the structures that persist continue to persist, and they lead to the existence of more persistent structures. Mother Nature has choices and she makes these choices based on the persistence of her past choices. Her “goal” is to generate more future choices with each tick of the universal clock, everywhere in both time and space. Life is one grand search. It is relentlessly searching for structures to build more structures. Per Darwin, all of these past choices are recorded in the fabric of her universe as a complex network of choices made. All past choices form the perpetual basis of future choices to be made. The output
for each round of choices made is used as the input for the next round. Local, regional and global fractal patterns are naturally created everywhere in time and space within the grand pattern of life. We should look carefully for them and pay close attention to them wherever they may be found.

A brief examination of known mechanisms of protein formation.

We know that life consists in large part of an intricate relationship between two general sets of molecules: nucleotides and proteins. We know that this relationship today involves a symmetrical system where proteins participate in the creation of nucleotides and nucleotides participate in the creation of proteins. We know that this is a beautifully efficient and effective relationship, yet we do not know specifically how or why this relationship came about or exactly how it works. We know that the genetic code somehow defines these relationships, but we’ve yet to elucidate the first principles of the molecular sets behind them. Science has, however, deftly uncovered many of the specific molecular mechanisms that execute the code in nature. It is useful then to briefly review some of those mechanisms.

Start by first assuming that molecular information already exists in nature and is stored in DNA – a huge initial assumption. This information exists in an idealized form we know as a double helix, yet specific proteins, called histones, wrap it into evermore-efficient structural configurations for efficient storage. More proteins then have the job of finding specific information and unpacking it so that it can be accessed in the execution of “the genetic code.” A specific DNA sequence is somehow selected by proteins, unpacked, and transcribed into a new sequence of messenger RNA (mRNA). Molecular information is now clearly being utilized and transformed by the system in a structural way. This sequence of mRNA must be delivered to a suitable environment where it can be translated into protein. The mRNA must first, however, be processed into a sequence somehow deemed suitable for translation before it is translated. The processing of mRNA produces many new structures that participate somehow in translation as structural agents. Again, proteins are involved in this processing and delivery mechanism, and there are apparently many possible options for getting this done. This is called transport and transcription editing, and it represents yet another form of molecular information from a purely structural perspective. These specific choices are made in tandem via a complex relationship between the mRNA and proteins.

Once mRNA is ready for translation, it is affixed by protein to a ribosome, or rRNA, which is itself a large complex of nucleotides and proteins. It is a given that, in order for translation to proceed, there are other complex populations of molecules available to perform the actual translation. Proteins must have already translated DNA into many independent transfer RNA (tRNA – should be translation RNA) structures. Other populations of proteins must also already
exist to pair existing amino acids with these tRNA. These highly conserved protein structures are known as aminoacyl–tRNA synthetases (ARS). The initiation and termination of translation itself requires still more proteins; however, once the translation is underway, we can only partially idealize it as a process of pairing codons with tRNA. The process of making an amino acid sequence is purely a process of pairing proteins with tRNA, not amino acids with codons. The process of making a peptide bond sequence is purely a process of pairing tRNA with mRNA and other tRNA.

Each tRNA contains at least one “identity sequence” known as an anticodon, and each tRNA consistently carries a specific amino acid given to it by a specific ARS protein. Although there are only four nucleotides that generally go into a codon, we know that more than four nucleotides participate in the formation of an anticodon; therefore, there are special rules for pairing the nucleotides in codons with the nucleotides in anticodons. There is a strong symmetry to the rules of pairing. There is even a special symmetry in these rules at the third position of every codon. Specifically, there exists an inherent ambiguity of nucleotide pairing between codons and anticodons at the third codon position. These special rules are known as wobble rules, because the tRNA can be idealized as “wobbling” in the act of precisely pairing itself with the appropriate third nucleotide position of a codon. The addition of extra nucleotides and wobble rules means that there are logically more possible anticodons than there are codons. This is clearly a form of molecular information. Life has proven it uses these wobble rules, in part, to decrease the number of actual anticodons, and, consequently, they also reduce the number of distinct tRNA molecules required in any specific translation. However, the set of ARS is remarkably always the exact same as the set of amino acids. So, wobble rules are related to information efficiency and structural translation of peptide bonds. After all, tRNA must now actually wobble. Regardless of specific form of information involved, the relationship between codons, anticodons and tRNA on a purely informative basis is one of natural expansion of molecular information. This is logically true simply because there are more molecular choices for anticodons and tRNA than there are for DNA, mRNA, codons, ARS and amino acids.

Just as amino acids and nucleotides are now perceived as whole molecular structures in our current model, so should whole tRNA be perceived in the context of protein translations. They are discrete molecular structures. Each distinct tRNA in the system should have a name just as each distinct nucleotide and amino acid has a name. These tRNA are huge molecular structures compared to individual nucleotides and amino acids, but their role is similar and their function is somehow related to expansion of molecular information in time and space. There is an alphabet of nucleotides, amino acids and tRNA, and the tRNA alphabet is the largest by far. Translation is a sequential process that must now include the sequential formation of tRNA “virtual polymers” in time and space as the direct translation of codon sequences. The sequence of translated
tRNA form sentences as an informative intermediate subset between nucleotide and amino acid sentences. These structural sequences immediately translate the structural information necessary to make any sequence of peptide bonds within any variation of the language of the genetic code.

Once the act of translation begins, it is minimally an act of creating a sequence of amino acids where there is known to be a consistent relationship between specific codons, anticodons, tRNA and amino acids. There is no empirical doubt about this, and this primarily forms the basis of our idealization of the genetic code today. This is the whole essence of the standard codon table that we use to symbolize, describe and understand our current idealization of this molecular code. However, the real translation is purely structural, and the tRNA structures cannot be ignored as information appears in full via total protein structure through time. After all, a single codon can have no meaning outside of a specific context because no sequence of amino acids can be formed without connecting each amino acid to another within any sequence. Nothing “means” anything in the molecular universe in the absence of all else. It is literally a molecular metaphor that draws structural meanings from structural comparisons through time.

Amino acids are known to be connected by peptide bonds, and peptide bond formation is known to not be a homogenous act with respect to molecular information. Some peptide bonds are known to form quickly while others more slowly. This is independent of the two amino acids involved, yet it is entirely dependent on the tRNA involved. In this way, tRNA are delivering molecular information in time and space to the formation of every peptide bond. This is a logical fact, not just idle speculation. Additionally, peptide bonds exist at all times in specific spatial conformations that ultimately define the overall structure of any amino acid sequence. They must, therefore, have a specific structure at the time of their formation. Remarkably, we currently have little knowledge of the “possible” number of conformations that can be delivered by tRNA to peptide bonds, yet this is another form of obvious molecular information that logically must be structural, and must be delivered by tRNA at the point of peptide bond formation. The information content of any amino acid sequence is always a subset of its sequence of peptide bonds. This is not a theory but a quite simple and logical fact of scientific reason. However, the logical implications or meaning of this fact is currently being universally ignored because the current model has no language or mechanism for handling it. It obviates all need to do so.

Virtually coincident with the formation of a primary sequence of peptide bonds, the nascent protein begins to form larger structural patterns that are typically referred to as secondary structures. These structures merely serve as fundamental building blocks toward still larger protein structures. The structures themselves are known to be remarkably consistent across the space of “all possible” protein structures, and they are known to consistently form patterns within and between secondary structures, which then accounts for “all possible” tertiary and quaternary protein structures. An “unfolded” protein is referred to
now as a random coil, but neither random nor coil is defined in this reference. Proteins are in fact structures built of coiled coils, but they are rarely random, and many different coils and ways of coiling them are possible. The genetic code helps define how this reliably gets done because the code itself represents an important form of molecular information.

Broadly speaking, there are four types of secondary structure: alpha helices, beta sheets, loops or turns, and extended segments of no consistent structure, or regions of structural “disorganization.” Although it is generally true that certain sequences of amino acids tend to always form the same structures, this is not universally true. Some sequences that are virtually identical are known to form different structures, yet some sequences that are entirely dissimilar are known to form the same structure. It is either true that sequence determines structure, or that sequence in conjunction with other factors form structure. The known evidence strongly supports the latter. Regardless, every protein structure is always defined by its sequence of peptide bonds, this is specifically what defines secondary structure and all structures above it, yet the same cannot be logically said of any sequence of amino acids. For any sequence of amino acids there are many “possible” sequences of peptide bonds because the amino acids are merely a molecular information subset of peptide bonds. On the other hand, every protein is always defined by its sequence of peptide bonds. Any ideological model that ignores this non-trivial point of information logic is naturally fraught with a risk of explanatory catastrophe. That’s exactly what we now have.

Once the initial sequence of peptide bonds is translated, the nascent protein enters a mélange of additional proteins. It also starts this process within the context of a specific structural environment, which is sometimes a specific structure called an endoplasmic reticulum (cool name, cool structure). These structures plus the environmental protein molecules shape, transport and edit the nascent protein toward its final destination. The original sequence is edited in several ways. First, it can be truncated, which basically means that a primary sequence of amino acids does not strictly determine even the primary sequence of amino acids. In addition to amino acid sequence alterations it also undergoes a series of backbone altering procedures that can impart major structural features to individual peptide bonds as well as global structures that might be formed by entire groups of peptide bonds. In other words, the protein begins to make coils and begins to coil the coils. The protein structure is guided via structural algorithms post translation in a complex way to its final structure by a set of proteins that can generally be called chaperones. In other words, a final protein structure might not be determined until additional molecular information can be added to the formation process. This additional information cannot, therefore, be entirely dependent on amino acid sequence. The structural information is derived not from amino acid sequence but from a specific molecular environment in which it forms. This is undeniably an important form of molecular information that must logically extend beyond any examination of “sequence-only” molecular information.
Once a particular protein structure is achieved, it is then required to interact within a particular molecular environment to form larger structures and more complex relationships. The character and quantity of this interactive behavior is entirely dependent on the entire set of molecules in which any particular protein molecule might exist. In one environment it might behave and interact in a particular way; whereas, in another environment it is likely to behave and interact in a distinctly different way. Therefore, counting the number of “possible ways” that a particular protein might exist and the probabilities of each way depends entirely on its specific environment. This greatly complicates the task of defining and quantifying any aspect of molecular information, but it is not ideologically acceptable to adopt the pure expedience of ignoring these difficulties.

Finally, all of the complex structural information that is generated by the languages that guide protein in its use of nucleotides to make more protein must then perform the function of making more nucleotides. It can do this by replicating old nucleotides, inventing new nucleotides, or translating old nucleotides into different forms of nucleotides, and then editing, shaping and deploying them as well. Regardless of the specific mechanisms, these functions are purely structural, and their languages must be efficient for relating the structural information in proteins to that of nucleotides. In other words, the information in proteins and nucleotides must be highly symmetrical so that it can be translated in both directions.

We can now appreciate the tremendous difficulty involved in modeling molecular information and the genetic code. There are too many sets, and too many functions to allow the language to be completely simplified. Different organisms contain decidedly different sets of DNA, RNA and protein. They have decidedly different macromolecular structures in general; therefore, they have multiple dialects for multiple languages that go into the global language of the genetic code. We can, however, begin to make progress in our modeling by noting the areas where life has conserved the most. These are sets in the system of sets that are the most biased and the most remarkably conserved. The set of nucleotides and the set of amino acids are two good examples. However, these sets cannot perform their functions in the absence of other important sets, like ARS and tRNA. Fortunately, we can begin to idealize all of these functions by recognizing that the most highly conserved elements of the system are found in the primary set of structures organizing the system. For all intents and purposes, we can idealize all of them either as dodecahedrons, or as parts of dodecahedrons at every scale of size and complexity. For instance, all tRNA within a single organism are more or less similar, and all of them can be idealized as the wedge of a dodecahedron carved from one edge to the middle. This means that amino acids can be further idealized as points for the larger structural parts – coils – of much larger parts of much larger dodecahedrons. Even if the super-structures do not end up making actual dodecahedrons, we can idealize them as symmetrical parts of highly symmetrical structures, and
therefore they become extremely efficient elements within larger sets, and functional operators in symmetrical functions. The functions that operate on these sets can become more efficient and reliable. More importantly, all of the information can remain so consistent that the amount of required information can be reduced dramatically. Most of the required information for forming, changing and repeating structures can be maintained in remarkably simple, compressed, symmetrical structures.

The Information Content of Molecular Languages

A single human hemoglobin molecule is statistically impossible. It consists of roughly 11,000 atoms arranged into 600 amino acids arranged into four chains more or less precisely located in time and space. Think of all the quarks! All of life's molecular behavior is animated by random molecular activity. Because it is random, the number of possible behaviors is infinite. The chances of this collection of atoms conspiring in random fashion to become a single hemoglobin molecule are zero. Yet, in the past second on earth, $2 \times 10^{21}$ atoms did precisely that. You, dear reader, depend on your body's ability to make $4 \times 10^{14}$ hemoglobin molecules every second of every day. How could such precise behavior ever be the result of totally random atomic motion?

Finding the answer must start with the realization that you and one second are not the appropriate contexts of time and space for understanding this natural phenomenon. The proper context for a single human hemoglobin molecule is all of space and all of time. The appropriate context for your hemoglobin molecules is one of your cells over your lifetime. The molecular information content of a single hemoglobin is virtually infinite within the context of an entire "random" universe for all time. This information content goes way down, however, within your cells because the information system is already organized around the molecular languages in your cells. This "pre-existing" molecular information allows your cells to process enormous amounts of molecular information reliably and efficiently. Not only did the translated molecules need to evolve, but the languages of translation need to evolve as well. In this light, each translation is a compressed form of evolution. Molecular translation is the evolution of structural information in the proper time and molecular context. Both the molecules and their specific languages are part of a specific molecular environment, and it is the environment that defines and quantifies the molecular information therein.

Realizing the unavoidable complexity of this system also must make us wonder whether we will ever be able to understand it. We know that we can partially idealize it somehow, but at what cost to our understanding? We must start by differentiating the sets of things to understand. Molecular information must exist on many complex levels. This appears to be the basic strategy taken by life in creating the system itself. In a perfectly random molecular world there are too many possibilities for molecules and events, so life gains control first by
greatly restricting the sets. It also keeps things highly consistent so as to minimize the number of logical relationships that must at some time be accounted. For instance, there are many possible amino acids, but life consistently uses only a few, and the ones it uses are logically related to each other. This is the first principle of organizing molecular information and its many languages of translation. The act of recognizing these few sets of nucleotides and amino acids can be accomplished with an equally consistent set of proteins. However, the act of structuring sequences of these amino acids is an inherently less consistent sequence of events. Therefore, life employs a fleet of larger structures – tRNA, mRNA and rRNA – to efficiently translate the minimum required structure into sequences of amino acids with the minimum number of sets, and a decreased number of elements in each set. Should life find that additional structures have become required, life merely expands the sets. All of the structures at all times must remain consistent; otherwise, the number of rules and amount of required information will increase exponentially. Life tries to economize, and symmetry is a great way to keep this internal expansion of primary sets from occurring.

All of the distinctly different sequences of DNA are quite literally distinctly different molecular structures. Life has the logical ability to operate on these differences in logical ways through space and time. A set of tRNA structures can logically and physically be idealized as efficient yet flexible instruments of information expansion in time and space. Although sets of tRNA vary greatly among different organisms, their basic structural logic does not. However, different sets of tRNA naturally lead to different dialects in languages of translation. These sets of tRNA cannot reasonably be excluded from our idealization of the genetic code, because they translate an important portion of its logic in time and space. The genetic code is not a single language but a highly consistent set of related languages. Its consistency is not a simple accident of descent, or a feature of some mythical last common ancestor. Rather it is a tightly organized, incredibly biased reflection of shared primary logic, or the unavoidable traits of our first common ancestor.

All of life’s sets of molecules and languages for logically relating them on earth are intensely biased. They are all tiny subsets of those that are logically possible, and they all remain incredibly symmetrical with each other. This is the exact same notion that led life to give domination of the system over to DNA, protein, the codon table, multi-cellular organisms and sexual reproduction on earth. They are all basically combinatoric strategies that allow the most information to be maintained in an efficient way, while still expanding the sets through time and space in an effective manner.
Expanding our view of molecular information.

Defining and modeling molecular information is an inherently complex task. I have no illusions that I am able to do it in any practical way, but I do firmly believe that we should now begin a rational process of idealizing it in a more useful way. I do not have the specific answers and methods required, but I can begin the general process of outlining and organizing the basic parts, and thereby we can begin along the difficult yet important path of defining and better understanding molecular information.

Before molecular information can be properly modeled, one must first recognize a fundamental difference between the information contained in any single molecule and the information contained in much larger molecular sets, or that is to say molecular arrays containing many interrelated organic molecules. This is especially important once one recognizes that the genetic code is a phenomenon of the latter and not the former. It is an organic molecular function operating in nature, and at a minimum it results in the synthesis of whole proteins not just amino acid sequences. The genetic code represents in reality the molecular algorithms involved in a complex evolution of molecular array information through time. The genetic code is, therefore, strictly a function of molecular arrays and not of individual molecules.

First admitting broadly the utility of any concept of molecular information allows us to then view single molecules as individual data objects. The basic properties of each object are defined by the atomic structure of a particular molecule given the specific probabilities of “all possible” structures for that molecule. There are many molecular structures for any molecular sequence, and they are all namable. Of course, composition is always a derived subset of sequence, and sequence is always a derived subset of structure. However, this hierarchical relationship is complicated further by the fact that molecules – especially organic molecules - must be placed into a specific context to ultimately derive their total information content. A single complex molecule, such as proteins and DNA molecules are, can visit an elegant ensemble of logically related structures during a given unit of time within a particular context. They also can change physical properties quickly depending on the specific environment in which the molecule exists. Heat and pH are two simple yet undeniably important examples of how environmental variables can play upon the complex task of elucidating total molecular information. Therefore, we can know that organic molecules are chemically, spatially and temporally dynamic data objects that must always be considered on the backdrop of time and circumstance.

Compared to a relatively simple idea of molecular information, the general concept of molecular array information is far more complex. The actual existence, quantity, character and “meaning” of individual molecular information is dependant on the total array of molecules in which any individual molecular data object exists at any particular moment in time. The situation is unavoidably
complex because each object actually contributes to the information of the array, yet the precise information of the object is strictly determined by the array itself. For instance, a specific protein molecule can be partially defined by its overall structure. This structure must always subsume its sequential identities of residues, or its linear amino acid composition, but it also interacts closely with its environment, including other protein subunits to become a quaternary structure, or a total protein. The totality of the molecular information of this single data object is derived from and dependent on the total environment in which it resides and the specific molecular array to which it belongs. If we take this individual protein molecule out of one environment and array, and place it into another, the quantity, quality and “meaning” of this molecular information can change dramatically.

This difficult concept can perhaps be grasped by imagining the transport of a protein from an extracellular to an intracellular environment. The shape, functions, and interactive capacities of the molecule are thereby demonstrably changed. The basic molecular information of this protein is changed in a quantitative and qualitative way. The illustration is more dramatic when we consider the case of a prion, a case where a single data object has the capacity to change the global information of an entire molecular array. This situation can lead to catastrophic disruption of the receiving molecular array, its information content and structure. Likewise, the simple addition or subtraction of heat from a system will change the information of an individual molecule as well as the information of an entire molecular array.

The case of messenger RNA provides us with another simple yet fundamentally different example of this general concept. Messenger RNA, like DNA, has the ability to combine portions of its structure with other portions of its own structure to create a distinct superstructure. So, included in the environment for an mRNA must be the mRNA itself. If the rest of the environment and array includes such molecules that an mRNA does or does not actually combine with itself, the molecular information of a single mRNA can be dramatically impacted by this single determination. The information of that same mRNA is also dependent on whether it is in or out of the cytoplasm, and therefore has or has not been given the capacity to “translate” information to other molecules in the array, or whether the heat or pH is relatively high or low so that translation is more or less probable. The actual molecular information for that mRNA molecule cannot be formally determined without knowing the exact context of the molecule, or more precisely without knowing the specifics of all the other data objects in the molecular array.

This is the recursive general context from which we must struggle to define and understand the genetic code. There is a distinct and profound form of complex relativity involved in molecular information of all kinds. Molecular information is relative to time, space and other molecules. The act of modeling and quantifying such a thing will be immensely complex. This lies in stark contrast to all prior efforts to model the genetic code. The classic model of the
genetic code is cherished for its simplicity, yet it utterly fails to capture the real complexity of the natural phenomena it purports to model. It is constructed on the three simple premises of co-linearity of molecular composition, directed flow of molecular information, and thermodynamic predetermination of protein structure. It is grossly inadequate as a definition of molecular information, as well as the supposed functions involved in its translation.

We can begin to appreciate the difference between this old model and a new expanded model by drawing a picture of the broad relationships in the new model. This less-simple picture is meant to replace the all-too-simple icon of the genetic code as dictated by the central dogma of molecular biology.

Figure 2.

Figure two is a cartoon or visual icon that serves as a general framework for understanding the basic structure of a more complex and robust natural molecular function that we might somehow metaphorically call the genetic code. It accepts the definition of the genetic code as a function that evolves discrete portions of a molecular array through time. Evolution is taken to mean the organization of information through time, and there are three levels of evolution that must be considered with respect to the genetic code. First is the evolution of discrete molecular information that results in the formation of a single protein. This is the standard, accepted, if not the completely inaccurate meaning today. Second is the evolution of the entire molecular array as this function is performed. Third is the evolution of the entire system of translation here on earth. All of these time scales or orders of molecular evolution should be taken
as enlightening toward our conceptualization and understanding of molecular information and the genetic code.

The code itself is a distinct entity like any individual molecule. It resides within and is made of a particular molecular array. It represents not a cipher but a time sequence of molecular events. It is an algorithm that molecular arrays can follow in time and space to achieve specific forms of molecular evolution. The code itself can and does evolve. It is universal in the sense that major features of the code can be detected across the grand diversity of arrays that perform it, but in its actual particulars it is far from universal. As the specifics of arrays change, so do the specifics of the actual algorithm.

Beyond the general concept of a molecular array, there are two main features that differentiate this kind of model from past models. However, because the past models are of molecular composition only - this model holds that structure always subsumes composition - this model subsumes past models completely. In other words, whatever could be done with past models can also still be done with this one, but so much more can be done with it as well. This is simply a more robust interpretation of molecular events through time. Two additional differentiating features of this basic model - beyond complex arrays of information - correct the demonstrably false ideas derived from the central dogma of molecular biology and the single target hypothesis of protein folding.

Starting with the central dogma of molecular biology, which says that information is contained in linear compositional sequence and transmitted stepwise to other linear compositional sequence, we can see that this is an utterly failed premise of molecular information. The visual icon of this dogma is usually viewed as follows:

\[
\text{DNA} \\
\uparrow \\
\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}
\]

Figure 3.

This is, in fact, not at all representative of the known empirical data. In reality, DNA does not in any way control the synthesis of protein or even DNA, and the information does not merely flow in one direction. A broader more enlightened view might reverse the molecular roles here. We now know that protein controls virtually all events within a molecular array. This includes the operations that form a protein sequence and a protein structure. It also includes the operations that form sequences and structures for DNA and RNA. Indeed, without protein, no DNA, RNA, or protein will ever be made. Protein gives information to the molecular array and derives information from the molecular
array. Merely because a portion of that information is stored – by protein – in the data objects of DNA does not give us license to draw absolute conclusions about the origin, function, quantity or flow of that information; however, this is, in fact, being done with the central dogma. To correct this historically false and overly-romanticized conceptualization we merely need to reverse the visual icon of the central dogma and add a few more symmetrical arrows of information flow. It now appears, predictably, that the central dogma gets the general relationships entirely backwards.

![Figure 4](image)

Figure 4.

The second profound distinguishing feature of this model involves the single target hypothesis of protein folding. This was erroneously derived from Christian Anfinsen’s Nobel Prize winning thermodynamic hypothesis of protein folding, which states that thermodynamic forces acting on protein composition will play a primary role in determining protein structure. This is true, thermodynamic forces do indeed play a very big role in determining protein structure, but they are not the only factors that determine protein structure. There are many empirically demonstrated factors other than primary sequence that play a role in protein structure, and they all somehow are derived from the environment and molecular array. Amino acid sequence is only one of them. We have yet to determine what all of these factors are, where, how and when they operate in the algorithm, but we have conclusively proven that other factors do exist and they are indeed a part of the genetic code as it should be properly understood. The fact that “silent” mutations can and do alter the folded structures of proteins is all we need know here. There are, on the other hand, no definitive studies that should allow us to accept the single target hypothesis. It is the dog that did not bark, and it can now easily be rejected.

As we finally move beyond the failed hypothesis of a single target of structure for any protein composition, we can quickly see that it greatly complicates any new model of the genetic code. Instead of a protein data object with merely a single possible structure, we must perceive that any protein indeed has many possible structures. This ultimately means that the quantity of information contained in any protein is a function of the number of possible structures for that composition. The structure can be formally defined by its sequence of peptide bonds, and the composition is merely a derivative thereof;
not the reverse as is commonly and erroneously concluded from the single target hypothesis.

An analogy for understanding this was provided me by Richard Merrick who used the classic game of twenty questions. The game is played by first imagining an object from the set of all possible objects. Another entity then begins a process of asking questions designed to algorithmically identify the object. It proceeds via a system of channeling the unknown object into categories or sets of objects. A complex process of logically juxtaposing and interrelating these sets inevitably and miraculously can lead to the identification of the unknown object. A new and informatively expanded model of the genetic code adopts a similar perspective on the process of forming a specific protein from the set of all possible protein. Instead of questions, the protein is channeled through time toward its eventual destination by a series of events that are dictated by the information contained in the molecular array. Each protein structure is but one of many possible structures, but that structure is reliably formed by the existence of other molecules in the array. For us to begin to track the molecular information in a single protein data object, we must at all times track its sequence of peptide bonds, or its time-dependent structure. At some point, each sequence becomes thermodynamically robust and therefore requires little to no further channeling.

From this broader perspective of molecular information and models thereof, we can see more clearly where the old model fails and how a new model can improve upon it. We can see that the premises of co-linearity, the central dogma, and the single target hypothesis define the old model completely. However, we can now also see them as an all-too-simplistic tautology: the genetic code is simple and linear because a protein can fold in only one way. A protein can fold in only one way because the genetic code is simple and linear. If this were empirically true then it would be logically sound, but it is empirically false so it has become logically meaningless, and all derived explanations are meaningless as well. This tautology has predictably failed us, so we must start the new process of modeling by rejecting it completely. The key elements of the old model must be discarded before the key elements of a new model can be installed and begin to make sense to us again.

The first step, after the rejection of the old model, is to begin the difficult process of defining molecular information. We have broadly done so here by admitting the relativity of molecular information, it is always relative to something much larger and more complex, which here is described as molecular array information. Furthermore, all of these concepts must be firmly grounded on the concept of evolution. Time is an essential component to understanding molecular information and its organization through time. The genetic code is an algorithm that evolves molecular information; therefore, it can never be fully understood outside the context of time on many levels.

As this new method of conceptualizing and modeling molecular organic information is finally debated and developed, we will recognize a pressing need
for a formal language to do so. Since the model involves pure information, which involve quantities and the interrelationship of quantities, the formal language of this model must take on a distinctly mathematical flavor. It will not be a simple language because it is not a simple system that we are modeling. Since the system involves fundamentals of space-filling and has been organized by clear principles of symmetry, we can imagine that the best language to describe it will share many features with geometry. Also, since symmetry is a hallmark of the organization of this system, we can expect that group theory and a language of molecular sets will prove useful as well. Since it is a code, cryptology will come in quite handy. Since the components of the system are highly interrelated and perpetually changing in subtle, complex ways, we might expect to also find elements of calculus that will eventually prove enlightening. Since our work is a human metaphor for a molecular metaphor, we can not accomplish this difficult task without heavy doses of linguistics and philosophy. The scientists of the future will perhaps focus more attention on math, linguistics and mineralogy than have the chemists and biologists of the past in their attempts to model the genetic code. At bottom, we still do not have enough command of the pertinent concepts to define a workable mathematical language today; however, we can anticipate that this crucial development of the model is in our future.

Despite the fact that we are not able to precisely define the exact components and their interrelationships within the system today, we are none-the-less able to make general observations about the information system, its origins, basic structures and patterns of operation. It is a self-organized system of molecular information that is founded on principles of complexity. The direction of evolution of this system is toward increased levels of complexity on all scales. Third order levels of evolution select for and perpetuate components of the system that can lead to evermore degrees of complexity. Symmetry of components in both time and space provides the best tool available for the system to accomplish this goal. A related observation is that information actually expands during evolution of the system at all time levels of evolution in this system. This is clearly true on the third order where the system can be seen to expand and increase in complexity and information quantity through time and throughout all life on earth. In other words, there is more organic information today - in the form of molecules and languages - than there was at the time of its origin, and this pattern of expansion can be seen to be increasing exponentially as the system continues to evolve. On the second order time scale we can see that molecular information expands as the molecular array performs the genetic code. Proteins are added to the array, which then merely participate in the processes occurring within the array to fulfill the basic expansion pattern. However, even on the first order level we can see that molecular information expands during protein translation. There is far more molecular information contained in an individual protein than is contained in a nucleotide sequence before it is translated. The genetic code merely imparts this robust set of molecular information upon the protein data object as the molecular information
evolves within the system. It is on these kinds of fundamental points that we can finally recognize the failure of the old model, as well as the need for and utility of a new model of molecular information and the genetic code.

Expanding our idealization of the genetic code.

As our concept of molecular information expands, so should our idealization of the genetic code. We only now have generalities, but countless specifics will eventually be required. Here are some commonsense steps toward finding those specifics:

- Locate molecular events in time.
- Define molecular events in space, including molecular sequence events.
- Fully integrate molecular information between nucleotides and proteins.
- Explain the desired function of the code, the various mechanisms to achieve that function, and the evolutionary path toward present day mechanisms.

On choosing a model to represent and understand the genetic code.

Information is real and exists in many forms. It is translated from one form to another by pure logic. Wherever information is translated, a language exists. Languages organize information by determining the nature of translation for that information, and so languages themselves become a form of information. The genetic code is real and it is a real language comprised entirely of molecules, but we have yet to identify this language in nature and perhaps never will. We do not understand the information it translates and we have yet to recognize the information represented by the genetic code itself.

The universe of molecular information is large, diverse and in constant motion. All of the molecules and forms of information must be taken relative to each other, which makes knowing this universe quite complex. We might never know the specific codes that dictate all of the motions, but we can know that logic is the gravity that binds them all together. Despite the vast diversity and constant motion, the code of life shares a common logic, and this we can know. And while there is a force of logic that is bringing molecular information together, so too is there a natural force driving things apart into more, and into more complex permutations. Life may use a common search method, but the results of the search guarantee an ever-increasing diversity in life.

The primary job of science is to make sense of the universe. Some of the explanations of the universe emanating from science today make sense and
some of them are purely senseless. The current model of the genetic code, based on an inherently illogical definition of molecular information, is generating explanations that are completely senseless. They are devoid of basic logic and reason, and people who argue them are merely being unreasonable. We must start over from scratch with a proper, logical understanding of our knowledge and system of knowledge. Only then can proper languages and models effectively expand that knowledge. A proper definition of molecular information and a proper model for the genetic code are the best possible places for us to start understanding life in this universe.

May the important debates finally now begin.