What is a Protein?

By Mark White, MD

Introduction

From Wikipedia we learn the following: "The word *protein* comes from the Greek word *πρώτα* ("prota"), meaning "of primary importance." Proteins were first described and named by the Swedish chemist Jöns Jakob Berzelius in 1838. However, the central role of proteins in living organisms was not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was a protein The first protein to be sequenced was insulin, by Frederick Sanger, who won the Nobel Prize for this achievement in 1958. The first protein structures to be solved included hemoglobin and myoglobin, by Max Perutz and Sir John Cowdery Kendrew, respectively, in 1958. The three-dimensional structures of both proteins were first determined by x-ray diffraction analysis; the structures of myoglobin and hemoglobin won the 1962 Nobel Prize in Chemistry for their discoverers."

So, is hemoglobin a protein? Apparently so. But hemoglobin consists of four amino acid chains, or four "protein molecules" that come together to make a "protein complex." Is hemoglobin one molecule or four molecules that constitutes one protein? Who gets to decide? Is hemoglobin one protein or four proteins that now constitutes a protein complex?

Most definitions focus on the fact that a protein is a "linear chain" of amino acids, so how does hemoglobin fit into that definition, since it is not exactly a single linear chain? Other definitions for protein focus on the genetic code, or the constituents that make proteins, but we can chemically alter the amino acids in a protein, and so we can easily go outside the genetic code to make molecules. Are they proteins? Regardless, since a protein merely represents a "sequence of amino acids," and this sequence always reliably folds thermodynamically into a single stable shape, it is the sequence and not the shape that defines a protein. That is easy to understand. However, it is long known that a protein can strangely become "misshappen," like in the case of prion disease, an example of which is mad cow disease. Do two molecules with the same amino acid sequences yet different shapes represent one protein or two different proteins? Who gets to decide?

What, precisely, is a protein? On what criteria do we accurately determine if a molecule is or is not properly called a protein? Do we base our definition on a single parameter or on multiple parameters? Which parameters do we use and which ones receive the most emphasis? Why does it even matter? Lest we forget: "The word *protein* comes from the Greek word *πρώτα* ("prota"), meaning "of primary importance,"

which is equally true in life and in language. The definitions for molecular information, molecular language and the genetic code all depend first on the definition of a protein. Strange, but true. Our exact definition of a protein determines the exact meanings of these other important words as well. So, again, what is a protein? The plain truth is that we have no workable definition of a protein at this time. There are profound flaws in all of the current definitions, and none of them have obtained consensus status. More importantly, we have no true understanding of what a protein actually is, so here we will try to find a definition that has real epistemic value. In order to do so we must seek a deep understanding of the kind of molecules that we are trying to define. Therefore, perhaps it is best if we begin with definitions of molecule, information and language. Further note that it is the genetic code that leads to the existence of proteins, so a definintion for proteins can never be divorced from a definition of the genetic code.

Information

In order to define information and language, we will introduce the simple notion of a set. A set is any collection of things considered as a unit. For instance, the faces of a cube constitute a set of six things. We are quite familiar with the six faces of a cube because we are familiar with ordinary dice. If we consider two cubes, one red and the other blue, for instance, each with numbered faces, then the set of all pairs of red and blue faces now has thirty-six members. Information is the collective relationships between each member of a set and the set itself. A numeric value for information is closely related to event probabilities. When information is quantified it is called entropy, which is more than a bit confusing, but we can easily put this simple conceptual notion of information entropy into the form of symbolic formula. The formulas for information entropy and thermodynamic entropy are highly similar. Again, from Wikipedia:

"The information entropy of a discrete random variable X, that can take on possible values $\{x_1...x_n\}$ is:

$$H(X) = \mathrm{E}(I(X)) \;\; = \;\; -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

where: $I(X)$ is the information content or self-information of X, which is itself a random variable; and $p(x_i) = \Pr(X=x_i)$ is the probability mass function of X."

This is a fancy way of saying that the information contained in one roll of a fair die is inversely proportional to the probability of each unique outcome, that is 1/6, and this has a value of 2.58 bits, and since the probability associated with unique outcomes in a roll of two fair dice is 1/36, the information is 5.17 bits. In other words, we know with total certainty that when we roll two dice that one of thirty-six possible combinations of faces will emerge from the roll, we just don't know which ones. Rolling the dice provides us with a quantifiable amount of information only if we know the relationship of the result to the members in the set of all possible results. The key then to quantifying

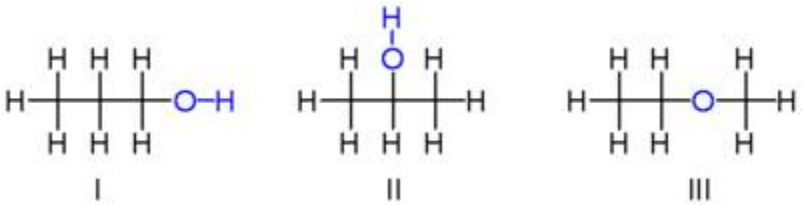
information is to first define the set from which something will be selected, and then define the relationship of each thing to the set as a whole.

Now that we have a concept of sets and information, we can define functions, codes and languages simply in terms of two sets. A function is merely a map between members of two sets. A code is a specific algorithm that describes a function. A language is a set of all logical relationships between two sets. In other words, a language can consist of many functions and therefore many codes. A language can be used to translate information between sets in various and complex ways.

Molecules

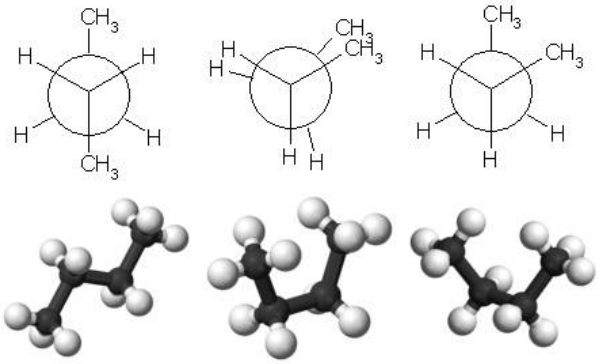
Now that we have these simple definitions, we can move on to the more difficult ones we seek, those of molecules, proteins and the genetic code. For these purposes, we will define a molecule as two or more bonded atoms. Of course, a bond can be many different things in many different circumstances. For instance, are the two strands of a DNA double helix bonded together or not? Are the four chains of hemoglobin bonded together or not? For our purposes, we will answer absolutely yes to both of these questions, and so we will say that a DNA double helix and a hemoglobin can each be thought of a single molecule. We will define a bond as a semi-permanent relationship between two or more atoms. We now have something of a circularity, however, where a molecule is two or more bonded atoms, and a bond is a semi-permanent relationship between atoms, so a molecule is now defined as a semi-permanent relationship between two or more atoms. In other words, whether a bond exists or not depends upon an entirely subjective reading of the term "semi-permanent," which then makes the definition of "a molecule" highly subjective as well. We confront the age-old battle between taxonomic lumpers and splitters. Just as the line between species is blurry, not sharp, so are the lines between proteins and molecules in general. One would think that something as fundamental as the definition of a molecule could be made with no subjectivity, but alas the universe is rarely that tidy, and so neither is human language.

It is common knowledge that the same set of atoms can be bonded together in many different ways, and each different way constitutes a distinctly different molecule.



I II III

Within each set of bonded atoms there is a set of bonding patterns, so this of course means that there is quantifiable information within the bonding pattern of a molecule. In other words, the identity and number of atoms obviously constitute molecular information, but the way they are bonded together constitutes additional molecular information. Perhaps less obvious is the fact that the exact "same" bonds can have many different spatial forms as well. These are known as conformational isomers.



There is now a set of spatial orientations within a set of bonding patterns, and that too represents yet another form of molecular information. So, we will define a sequence as a one-dimensional delineation of information, and we will contrast that with a definition of structure as spatial information. In other words, structure cannot be one-dimensional, but sequence must always somehow be a subset of structure, and molecular information is always structure. When it comes to molecules, sequence is only shorthand for structure, but sequence should never be confused with structure. So, molecular information consists at a mininimum of the atoms involved, the bonds between them, and the precise spatial arrangement of all the atoms in the molecule. However, this still is not enough to fully define molecular information, and we shall see why this is true as we take a quick look at one of the oldest and best known proteins, a molecule known as human hemoglobin. We will now discover that molecular information must include time as well as space.

Hemoglobin is a common protein that is synthesized in bone marrow by a blood cell called an erythroblast, and it is then packaged into red blood cells. In many different respects, hemoglobin can exist in several different forms. First, note that each hemoglobin molecule is actually a composite of four amino acid chains of two different types – alpha and non-alpha. There are different forms of each chain. The four chains come together to create a specific spatial structure that carries four iron molecules, and this arrangement is well suited for carrying oxygen in blood. The overall structure of hemoglobin changes considerably depending upon the number of oxygen atoms it carries. The function of hemoglobin depends on the fact that it can rapidly assume different forms in time and space. There are two alpha chains and each has 141 amino acids. There are two non-alpha chains and each has 146 amino acids. Assuming that the average amino acid has roughly 19 atoms of varying amounts of carbon, hydrogen, nitrogen, oxygen and sulfur, there are roughly eleven thousand atoms in a single hemoglobin molecule. When one considers that each of these atoms must have a specific relationship in terms of bonds and spatial orientation on a timeline relative to every other atom in the molecule, one must consider that there is an enormous amount of molecular information in a single hemoglobin molecule. Note also that the typical human body produces approximately 4 X 10^{14} hemoglobin molecules per second. Your body, for instance, is thus busily arranging 4 X 10^{18} atoms precisely in space every second of every minute of every hour, and this is merely to supply the proper hemoglobin information, which is only one of tens

of thousands of proteins in your body. What exactly is this molecular information, where does it come from, and how can we ever hope to understand it?

The formation of hemoglobin depends on the thermodynamic activity of atoms on many different levels. However, thermodynamics alone will never allow us to understand hemoglobin. The probability of these 11,000 atoms - doing just *once* what they must do so often – the probability of this atomic behavior alone is so low as to be considered zero. The answer, of course, is that hemoglobin is part of a much larger atomic system, and the collective atomic march of molecular information has been systematically moving up hill against random atomic behavior for over four billion years. We can never consider atoms in isolation. And, more importantly, our understanding of this entire system requires a complete rethinking of our notion of random. The information that we find in hemoglobin was not put there instantly; it has been slowly accumulating for billions of years. We just need to learn how to recognize it. Context is everything. Information is context. Unfortunately, our shabby current definition of a protein does not allow us to understand the system that makes protein. A protein is more than a sequence of amino acids. Either you "get it" or you don't. But in order to begin to understand this critical basic concept, we must first break down the protein system into the component parts.

In order to properly understand protein, one must first know that it is a dynamic molecule. In other words, proteins are not static; they move through time. Proteins are time-dependent molecules, and they depend on time at three distinct levels. There are three separate orders of time-dependent molecular information within every single protein. The first-order of time-dependent molecular information within a protein equates to its ontogeny. Every protein has a birth and a death, everything inbetween is a protein's ontogeny. No protein can exist without a personal history, so to speak. Each one is born of an interaction in time and space of many other molecules, and each one exists as an interaction in time and space with many other molecules. Proteins are not molecular snapshots they are molecular movies.

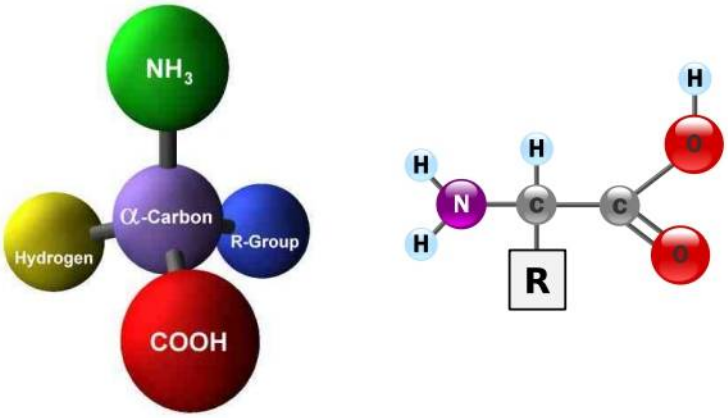
The second-order of time-dependent molecular information within a protein equates to its phylogeny. Every protein has a family tree, and most every protein will lead to many descendents. If we consider the space of all possible proteins, we can see that this represents a search space, and life systematically moves about it. Each protein is a product of that search, and so each one has a search history, and each one can be related to others by their search history as well. The molecular information within human hemoglobin was not created for human hemoglobin, per se, but was fortuitously found within the path of a much larger search.

The third-order of time-dependent molecular information within a protein equates to the overall protein system. If we consider today to be $Time_1$, then we can imagine a $Time_0$ when no system for making proteins existed on earth. Between $Time_0$ and $Time_1$ we know that a tremendous amount of molecular information has accumulated within the system itself. We also know that the overwelming majority of the molecular information contained in human hemoglobin is exactly this kind of molecular information; it is third-order time-dependent molecular information. In other words, before a search of protein space can be conducted, a protein search space must be created. The vast majority of protein information on earth today is derived from the search space upon which it

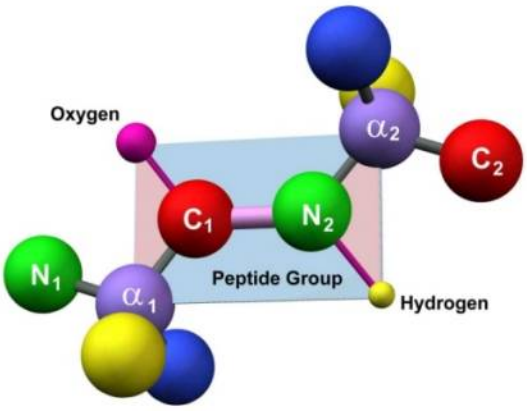
operates. Therefore, in order to understand the game being played by any individual protein, we must first understand the game being played by all proteins together.

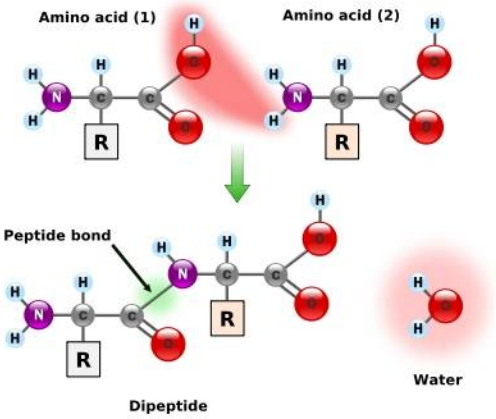
The Protein System

The atomic system that builds proteins has many parts, but the most basic set of parts is the set known as amino acids. There are twenty "standard" amino acids in the protein system today. What is an amino acid? An amino acid is simply a molecular tetrahedron. It consists of a central carbon, known as the alpha carbon, and carbon has a valence of four, so we can think of it as a tetrahedron.



An amino acid conceptually is an extended carbon, or somewhat of a carbon Swiss army knife, where there are four standard attachments. The first attachment is merely hydrogen. The second attachment is a carboxyl group, and the third is an amino group. The fourth attachment is the variable utility part of the carbon knife, it is called the R-group. Each amino acid is defined by its unique R-group, so there are twenty standard R-groups within the set of amino acids in this particular protein system. The amino group and the carboxyl group are the hook and loop that allow a chain of amino acids to consistently form. The bond between them is known as a peptide bond.





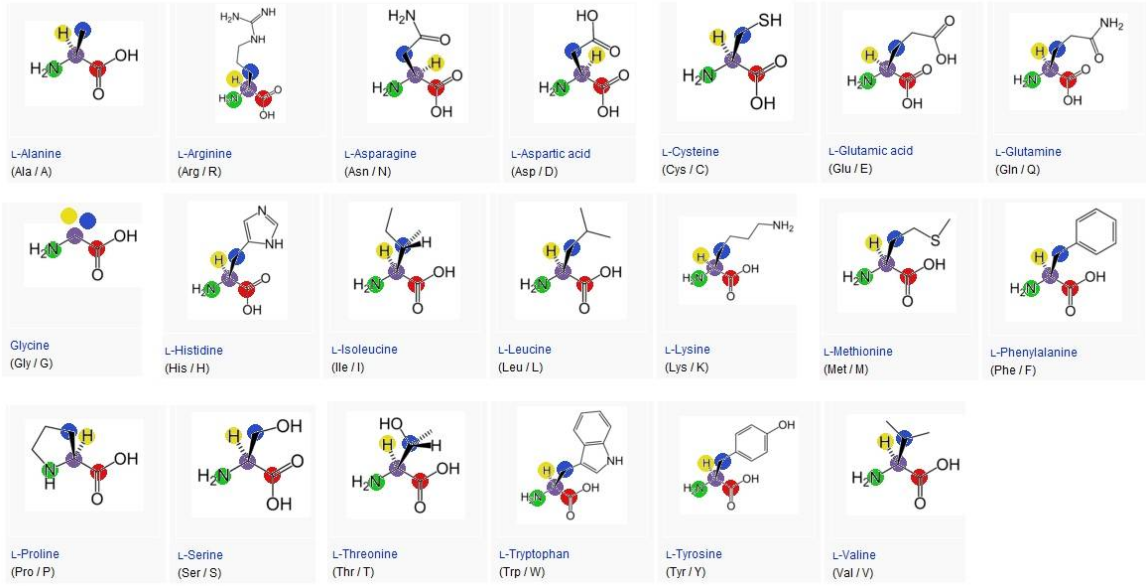
Within the sequence of amino acids of any protein there is very little information to be found in the specific atomic partners of this bond. The bonding partners are always the carboxyl carbon and the amino nitrogen, but there is considerable information still to be found in their overall spatial conformation. Molecular information is always structure, and there is quite a bit of structure in every peptide bond. They can exist in a cis or trans conformation, and the alpha carbons attached to each of them can rotate in 360 degrees. The entire chain is dynamic, and so over the lifetime of any one chain each peptide group might visit many different members of its set of possible spatial conformations. Depending on the exact R-groups involved in a particular peptide group, there is a set of conformational isomers. This of course is molecular information. The set of all peptide group conformational isomers for an entire chain represents a combinatoric group of chain conformational isomers that is logically quite large, and so there is also molecular information contained in the shape of every chain. This kind of molecular information can be seen as first-order molecular information, but keep in mind that it is mostly inherited from the third-order molecular information that has accumulated over billions of years. In other words, the nature of the set of standard amino acids determines the nature of the sets of conformational isomers for all the chains they make. This information is currently ignored, as if it were "free" information to the system, but it clearly is not.

At the point of formation, each peptide bond must have a specific conformation. This is not to say that we know the conformation, or even that we know it to be consistent or predictable, or to say that it even matters at a later time in the ontogeny of any particular chain, but we can say with absolute certainty that every peptide bond has a specific conformation at its point of formation. Each growing chain then must also have a specific conformation precisely when each bond is made. Therefore, the formation of each peptide bond signifies the formation of a specific molecular structure that includes all molecules involved for that instant in time. Every amino acid chain formation then actually represents a sequential formation of many molecular structures in time and space. The number of possibilities in time and space for conformational isomers at the point of formation of each bond is truly enormous, and of course they each then change over time according to a yet unknown set of physical parameters. However, each bond

has its own birth, death and ontogeny. The set of bond ontogenies for a chain determines the entire chain ontogeny. It's just that simple… but not simpler.

By the way, where is molecular information?

A larger point about molecular information has apparently been missed here as well. First, consider the set of all possible atomic building systems. It is quite large, yet this particular system has settled on carbon as its base. Now consider all of the ways that carbon might be used within an atomic building system, yet this particular system has settled on a set of carbons with four bonding partners. Now consider the set of all carbon atoms with four bonding partners. How large is this set? How many different potential bonding partners are there, and how many different ways can carbon have four bonding partners? This particular system has settled on the amino acid as its base building unit. Now consider the power set of amino acids, or the set of all subsets of amino acids. This set is inconceivably large, so the selection of one subset from all possible sets represents a truly staggering amount of molecular information. This particular protein system has selected just such a set. It is not an arbitrary set. The set is not "free;" it had to be found. Selection is work and selection is information, and this set represents a huge amount of work and information. This particular molecular set was not given to the system before the first protein was made. This is a set selected by the system over long periods of time, a set selected specifically to make all proteins within the system today. This set defines the current protein search space, and the protein search space defines all proteins. In other words, every protein made by this set of molecules inherits a tremendous amount of molecular information from the set of standard amino acids. The set itself represents a tremendous amount of molecular information.



Think of the standard set of amino acids as a choice made from a huge set of possible choices. This, of course, represents molecular information, and so now every protein made by the system inherits the information of this set. Every protein within the system inherits the information of the system itself. Compare this to a human alphabet or a set of numerals. For instance, we have mostly chosen to use a number system with ten numerals. We then chose to use a sequential power system to represent numbers larger than ten. So when we see a sequence of numerals, such as 7592413, we automatically know to multiply the one on the far left by 10^6, or we know that this numeral represents the number seven million and not just seven. In other words, the entire system imparts information on the sequence. The information is not just found in the sequence itself; it is found in the sequence within the context of the system. Note, this particular sentence is merely a sequence of letters, but the information within it is derived simultaneously from the sequence and the English language. The language organically grew up around the nature of sequential sets of a primary set of symbols. Nobody planned English. The parameters of the system are dictated by the nature of the sets of parts. The protein system is fundamentally no different, except it is more logical, and therefore more efficient and more informative.

Compare and contrast the alphabet, the ten numerals and the standard set of amino acids. Consider the information within each set member and the information of the set as a whole. What is the nature of the information within each set? Nobody can really know for sure, but the set of amino acids is obviously composed purely of molecular information, and each member obviously contains a substantial amount of it. Each molecule inherits the information of its atoms. Collectively, the atoms in each amino acid represent a certain amount of molecular information. Atoms have choices, and those choices have been made to become amino acids. But when sets are used to make sequences of their members, what is the nature of the information in those sequences? In the case of atoms and molecules it is, again, purely structural. Sequence is always a tiny subset of structure, a shorthand for partial human understanding. But, when those sequences interact with other sequences, as they inevitably will, what is the nature of the information in those interactions? Again, obviously, it is purely structural. So, how many additional sets are made from this particular set of amino acids? What is the nature of the information contained in all of these additional sets? How complex can we expect this information to become with time? Obviously, it is all molecular information, and so it too is structural, but absent any specific and useful, perhaps qualitative if not exactly quantitative answers to these extremely difficult questions, I don't think it is too terribly hard to finally realize that the set of standard amino acids is perhaps the most informative set known to man. Name a close second.

It is a crying shame that when we think of amino acids we hardly consider the molecular information that is contained within each one, let alone the information that is contained in the entire set. We foolishly think of them as whole units, not as collections of parts, because the behavior of the parts is so consistent. They might as well be spheres instead of rich molecular configurations. But consider this, all of the proteins made within the system are in some way a product of this basic molecular set. It is the uniqueness of this set of twenty molecules that dictates the uniqueness of every protein. In other words, all proteins inherit the information that is at first contained in this set of molecules. We can easily imagine a handful of additional amino acids that we might add

to this set, and these additions would open up vast new spaces of proteins for life to search, to be sure, yet life sticks to this relatively small and remarkably consistent collection of molecules. Why?

The consistency of this set begins at the third-order of molecular information, but it obviously extends through the second and first-order as well. The answer to the question of why all of the trillions of trillions of cells of life on earth stick to this particular molecular set is because this is consistently the best set. It is the best set for consistency. There is a tremendous information value in consistency, and this set consistently transfers molecular information throughout all levels of the system. And it is only by maintaining consistency at the third-order that consistency can be achieved in second and first-order information as well. The consistency of third-order information allows for consistency in second-order searches of protein space. It is obvious that information is exchanged vertically and laterally everywhere and at all times within the protein search. Sexual reproduction is the best way to transfer information both vertically and laterally, and so sexual reproduction is the standard method used for accelerating the search. Viruses also provide an additional mechanism to achieve a lateral transfer of second-order information. Note, however, that with this system there are also many sequence transformations that are inevitable. The system welcomes them, seeks them out, actually. However, if consistency is absent at the third-order, this type of second-order activity becomes impractical and inefficient.

For instance, human hemoglobin's distant ancestors surely had nothing to do with the function of oxygen carrying. Perhaps the function of each hemoglobin chain was completely unrelated, but this particular combination of chains was eventually found to make a good fit with each other and a good solution for the distantly unforeseen function of oxygen carrying. It is even remotely possible that one or both of its chains first appeared in the search as "backward hemoglobin." Perhaps the sequence of amino acids was entirely different, and perhaps all of the amino acids at one time changed all at once. Such a thing might be thought to be a "random" catastrophe for hemoglobin's distant ancestor, but not within an integrated system of information leverage such as this, This kind of serendipity is only made possible by the logic of a highly integrated system of information that operates on many complex levels. However, it is only from the consistent, integrated and highly optimized arrangement of this particular set of amino acids that such things allow the information in one chain to be shifted, inverted, mirrored, transformed in every way imaginable, and yet it still produces a complex molecule that is much better than random within this protein system. The concept of random has now been completely stood upon its head by this kind of self-reference. Note, all of the molecules are selected within the system for the system. In other words, there is no such thing as a truly "random" protein within this system as it operates today. All proteins are logically related to all others. A truly random molecule stands no chance of survival within this system.

It is the nature of the system and the search itself that is so consistent, and that is why we see so much consistency from the molecules within the system. In other words, the informative value of a "good" set of amino acids can be appreciated at every level of the system. The value of any set of amino acids is difficult to ever fully appreciate, more so if one completely lacks an understanding of the system that employs them. Today's description of this system as "one-dimensional" is perhaps the most grotesque distortion

of reality since the ideas of Karl Marx. His ideas have never worked in practice, and neither will these. Surely the one-dimensional genetic code stands as the worst idea in the history of science. Name a close second.

Perhaps the most remarkable feature of this particular set of amino acids is its ability to provide phenomenal consistency in the form of first-order molecular information. In other words, the complex molecules it generates are counted upon to have consistent ontogonies. How can any molecule of such complexity ever be counted upon to have a consistent ontogony? How could the life of any complex molecule subjected to the persistent ravages of thermodynamic collisions ever become so predictable? Yet they not only appear to be thermodynamically well-behaved, they are also well-suited to interact with all of the other molecules they encounter during their lifetimes. How does one persuade any molecule to behave, let alone persuade all molecules to behave together? Why does purposeless atomic behavior emerge in a form that appears to us to be so predictable and purposeful? After all, there is nothing written in the fabric of the universe to say that two chains must combine with themselves and two others to become hemoglobin. Is there? But this kind of molecular information is reliably produced by this protein system with remarkable consistency. How? The answer is that protein behavior is purposeful in the same way that sodium and chloride find purposeful behavior in the structure of salt. The only difference is that the structures of life are infinitely more complex than the structures of salt. Obviously this kind of molecular behavior is an extension of the set of amino acids that was selected by the system. A different set of amino acids, in all probability, would not have this uncanny ability. The information inhereted by protein from its set of amino acids manifests itself in the form of protein-protein interactions. The proteins inherit this ability from the system itself. Consistency of molecular behavior is not an accident, it is a function of the system. It is a system of consistent molecular behavior. The consistency within the system is a function of molecular information. Molecular information is multi-dimensional, and must always be considered within the context of time. After all, "molecular behavior" is merely a manifestation of atoms moving in space through time.

Imagine, if you can, the process that makes a single hemoglobin molecule, and imagine it from an atomic perspective. Let's just pick up the action from the point of the second amino acid being added to the first chain. For 146 amino acids to be bonded together we must first have 438 nucleotides in a messenger RNA to direct the process. Let's assume that there are 25 atoms on average in a nucleotide. This means that for the 2,700 atoms in this one emerging amino acid chain to find their proper bonding partners, we must have roughly 11,000 atoms in mRNA occupying their appropriate portions of time and space. Ignoring the critical roles of rRNA and aminoacyl synthetase, we now merely need wait for the "right" tRNA to randomly tumble by. Assuming that each tRNA is, on average, composed of, say 2,000 atoms, we merely need count on the "proper" behavior of roughly 300,000 atoms in various tRNA to produce the first of four chains to make hemoglobin. In other words, it will take well over a million atoms in tRNA to ever hope to locate the atoms of one hemoglobin, and this complex atomic dance must come off without a hitch every single time. Why don't more of these atoms get it wrong, or why aren't there groups of "bad" atoms that consistently spoil the party? Nonetheless, once all four chains are made, they miraculously join up to make the 11,000 atoms of hemoglobin, perfectly arranged in time and space. What organized all of these

atoms in the first place? How can purely random molecular behavior become so consistent? By the way, why does this system so closely resemble our systems of numbers and letters? Nothing could be more different than an atom and a human, so why do these languages that they each are using look so freaking similar?

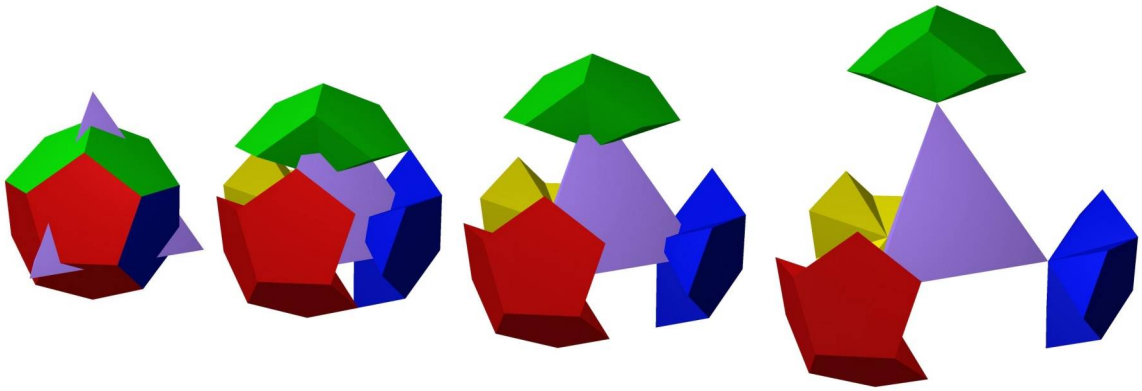
The answer to this seemingly unanswerable puzzle is that this is truly a molecular information system. It is nothing less. After all, we can logically know that all information systems must achieve common goals, and so all of them can be expected to find common mechanisms to achieve those goals. What are their goals? At bottom, every information system must transfer information between sets, and so those sets must be logically related. There is no way around it, but there are good ways and bad ways of organizing information between sets. Note that the relationships between sets allow languages to organically grow up around them. This is what languages are. And as languages grow, sequences naturally form that can be compounded and self-referenced. This is how languages grow. The languages then act as information generators between the sets. In other words, information accumulates within each set as a result of the simultaneous accumulation of logical relationships between the growing sets through time. It is a system of self-reference for making more self-reference. What could be more precious?

English, math, molecules, they all serve the same purposes at bottom. Perhaps the most profound, efficient and informative language on earth is the language of molecules. One might even venture to say that all other languages are derivative of it. Is there a plausible argument against this? It stands to reason then that if one could somehow unlock the fundamental secrets of this one language, one would have a leg up on all languages. What could be more important? Unfortunately, we are miles away. We are barking loudly, but alas it is up the wrong tree. Mankind has obviously suffered a serious setback as the result of an obvious scientific hoax, the hoax of the genetic code. We will perhaps eventually recover, but how long and at what price?

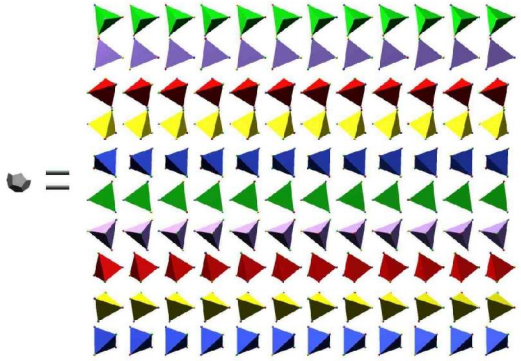
Think of the human system of mathematics. On what is it based? It is at first based on the existence of sets of numbers. What is a number? Nobody really knows, but whatever it is, there are sets of them, and there are logical relationships between these sets. Obviously, man has already discovered several of these relationships and so man has developed all manner of ingenious symbols to communicate this, and languages to translate information from one set to another. Furthermore, there are now languages to translate the ideas about these relationships between one brain and another. Math is big business, as far as man is concerned. It is all so gloriously self-referential. Note how the language of mathematics has organically grown up around the logical relationships between sets of numbers. The more we discover about mathematics, the more we discover that there is to discover about mathematics. It is a self-perpetuating process with no logical end in sight.

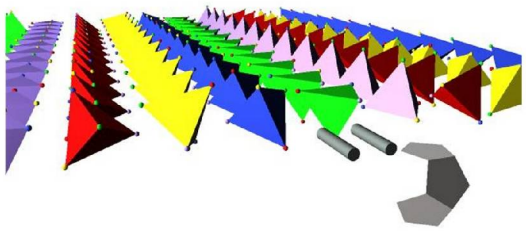
Now consider molecules. The situation with molecules is virtually the same as the situation with numbers. There exist sets, and there exist logical relationships between these sets. Through time, a language organically grows up around these relationships, and through time the amount of information grows exponentially. The only question now is, where did these sets get started? In other words, on what logical relationship could any molecular language begin the critical bootstrapping process? The answer surely must be structure, because all molecular information, at bottom, is structure. But on what

structure can we begin a language? The answer is obvious – there is no structure on which a language can begin. Duh. A language will require two structures and a logical relationship between them. Okay, smartass, on what two structures can a molecular language begin? Again, the answer is obvious. We must search all possible structures and find the most robust relationship between any two of them. Our choices are frightfully limited, so it is fortunate that we can find a fabulous relationship within the pool of limited choices. There is a suitable relationship between a dodecahedron and a tetrahedron, and it is identical to the relationship between nucleotides and amino acids. What are the odds?

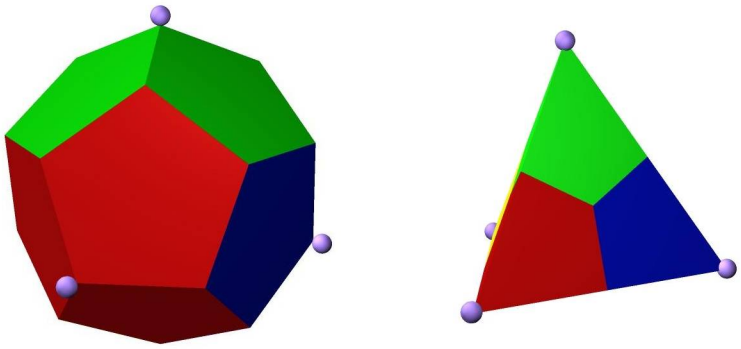


In fact, for the set of all conformational isomers of a tetrahedron a single dodecahdron can contain exactly 120 unique tetrahedrons.

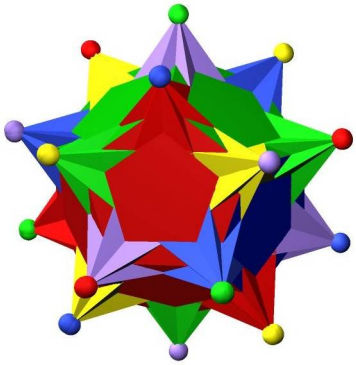




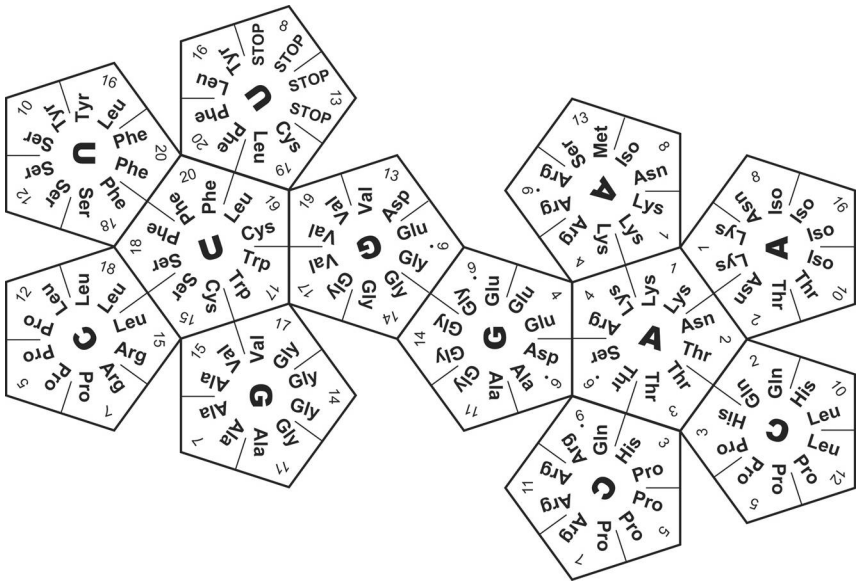
Moreover, simply by using a system of four colors, a sequence of any three colors can be used to specify any one of those 120 tetrahedrons.



So, an entire language exists between these two basic structures even before a single molecule ever randomly tumbles by to take advantage of it.



Now, my question to you is this: What are the odds, then, that sets of molecules would perfectly emulate this structural language, and yet there could be no logical relationship between the language of structures and the language of molecules? I would say that the odds are so small as to be zero. In other words, there is logically no chance that the genetic code is not a structural language at the very bottom.



This represents two different maps between sets. First, it demonstrates a map between the set of intersecting planes in a dodecahedron and the set of tetrahedral conformational isomers in a dodecahedron. Second, it represents a map of the set of nucleotide triplet permutations and the set of standard amino acids. The maps are isomorphs, and they each, as it happily turns out, represent symmetry groups. What are the odds? However, neither of these are languages, but languages can be organically built up around them. Do not get confused between maps and languages. A map of the earth is no more the earth than a codon table is the genetic code.

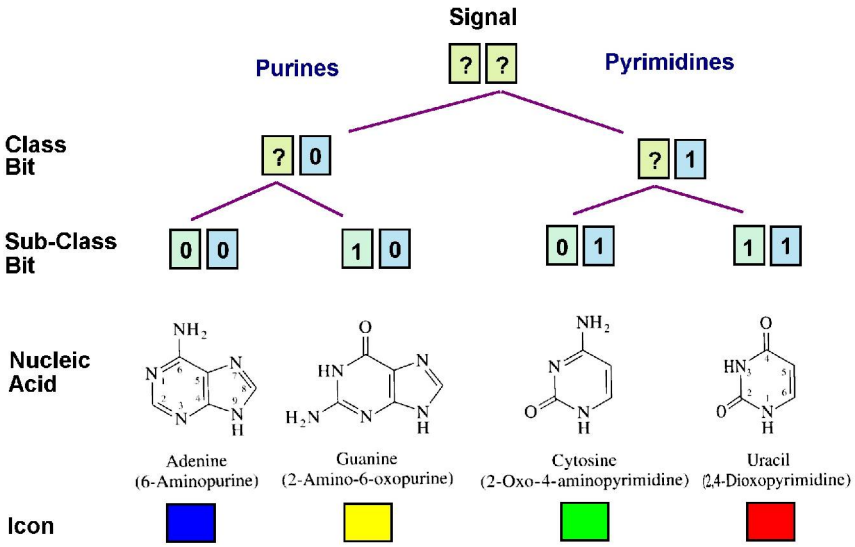
Traditionally, the genetic code has been defined as the map between nucleotides and amino acids, or what is known as the codon table. Logically, this means that a protein is defined as a sequence of amino acids. We now know that there is no epistemic value in either of these definitions. They are frauds. A protein is not merely a sequence of amino acids and the genetic code is not merely a map of codons. We are still a long way from understanding either one of them, but at least now we can finally realize that we do not understand either one of them. We do not even have suitable definitions for either one of them. We can, however, say that the genetic code translates molecular information. Molecular information is always structure. So, the genetic code is based on structure.

The consistency of the system is based entirely on the consistency of structure. Logically, without that, there is absolutely no hope that a consistent system will somehow appear across all of life and remain remarkably consistent for over four billion years. At $Time_0$ there were no complex molecules, but there was a language of structure upon which they could bootstrap an entire system for building molecules. All languages must start somewhere, and this language obviously started right here, in the relationship between a dodecahedron and a tetrahedron. It has grown into a complex language indeed, but every atom that participates must have a way to understand the language. Structure is the only thing that any atom ever understands. Unlike humans, atoms are stupid. Strike that. Unlike humans, atoms merely "understand" the core logic of the languages they use.

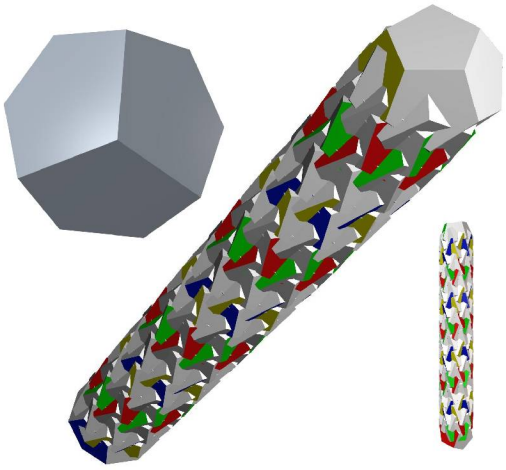
The other set

We have discussed proteins, molecular information, languages and the genetic code, but there must be two sets before there can be a language between them. The other set in this case is compossed of molecules that use building blocks called nucleotides or nucleic acids. Unfortunately, there is no convenient general term, like "protein" to describe the entire class of macromolecules they make. There are two basic types, DNA and RNA, and many important sub-types, so we will coin a new term to balance them with proteins; we will call them nuteins. The genetic code can then finally be seen as the set of logical relationships between the set of nuteins and the set of proteins. It is not merely a protein construction system, it is a protein-nutein growth system. The genetic code is not a codon table, it is not merely a map between codons and amino acids. There are huge numbers of molecules involved – there are even many important molecules between codons and amino acids – and there is a huge amount of molecular information in play beyond "linear" sequences. Remember, molecular information is not sequence. Molecular information is structure. Proteins can fold in more than one way. Molecular information is always structure plus time. Proteins and nuteins are not sequences except that they obviously can be cartooned as sequences of structure. The sequences are the molecules way of managing time with structure. There is a big difference. Each codon is a structure and each amino acid is a structure. These two sets of molecules are not the entire language but merely the primary building blocks of each set of larger molecules. The actual language, which has been improperly named the genetic "code" is a complex, robust, efficient and eminently interesting language. Like English or C++, or Calculus, it is a full language that cannot be learned completely in five minutes. In fact, it cannot be completely learned today at all, because our understanding of it has been so grossly distorted by a demonstrably false belief in a one-dimensional icon. The genetic code at a minimum includes DNA, mRNA, rRNA, tRNA and protein. The molecular information that it transfers between these sets is complex, multi-dimensional, and time-dependent on many different levels. All of the information is molecular information and all molecular information is first based on structure. The set of nuteins merely inherits a dodecahedral structure and the set of proteins inherits a tetrahedral structure. All other structures can be built from these sets, but this is where the sets are founded. The dodecahedron is to the tetrahedron as nuteins are to proteins.

We will focus here mostly on DNA, but please know that translation involves RNA, which has a different set of nucleotides, a much more robust set of nucleotides. We can interchangeably use the symbols 'T' and 'U', 'T' for the DNA base thymine, but preferring 'U' for the RNA base in RNA. Generally, DNA is more interesting to us in this particular context because it was chosen for this language in nature as a basic molecular set, one that is remarkably well-suited for consistency. It consistently makes perfect structures. It also perfectly sends two bits of information with every base in a chain.



We can now crudely imagine that each base represents one opposing face on a dodecahedron, and the double helix of DNA crudely represents a crystal that is composed of a sequence of dodecahedrons.



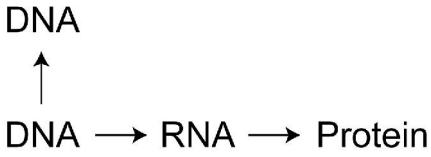
The codon table can now crudely be seen as translating dodecahedrons into tetrahedrons. What could be more precious? Note, however, that information is stored in DNA in a tightly compressed format. How else should vital information be stored? We can only begin to imagine, however, the exact logic of this compression scheme. It is far beyond the grasp of the strongest human minds with the most advanced digital technology. However, the puzzle of how we can store structures efficiently in sequences is solved by realizing that we store them in sequences of perfect structures. There are many levels of information efficiency involved in this scheme that we have been painfully missing by focusing so intently on codon tables instead of on the codons themselves. We must now begin to realize that the information contained in one sequence of DNA comes in many different forms. Start with the notion that each

sequence is actually a unique structure, but don't stop there. For instance, each sequence also actually represents many sequences, starting with its complement and extending to its inversion, frame shifts and all potential point mutations. The symmetrical structure of the codon map pattern merely reflects the symmetrical structure of DNA itself. These particular molecules were selected for this four-molecule set because of their ability to form binary pairs to be used in triplet permutations. In other words, the specific molecules we find today did not select one codon map from many; the codon map selected these sets of molecules from an almost infinite set of possibilities. The language selects the molecules, the molecules execute the language. First there is a structural language and then there are sets of molecules to build upon it. Molecular information proliferates in the process.

Think about all the information available to humans today. Think about how much information there would be in the absence of language. Molecules are no different. Language creates information, and the information manifests itself within the language.

Dogma dies hard

Perhaps one of the most bizarre, persistent and irritating oddities of the scientific fiasco that is the one-dimensional genetic code is the idea known as the central dogma of molecular biology. This dogma has been stated, re-stated, modified, apologized and enshrined in numerous and creative ways. We can say it simply as this: DNA makes RNA makes protein, and DNA makes DNA. We can even draw a picture of it.

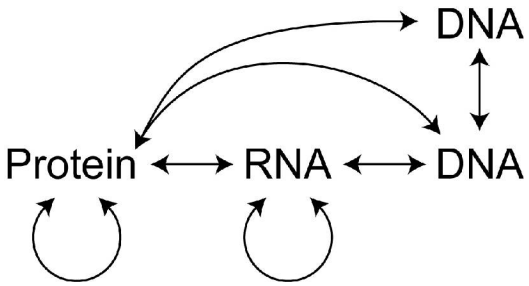


This illustrates the supposed "flow" of "information" within this all-too-simple concept of a molecular information system. This is nothing more than a cartoon of a bad joke. It should serve as a painful, yet graphic reminder that our understanding of proteins and molecular information has been woefully deficient for many decades. We need to realize that molecular information, in order to exist at all, must flow freely in both directions between both sets. However, once we get a proper intuition about molecular information we can easily correct the error and draw an improved picture.

First, think of an ant colony. Information about the colony is complex, it exists mainly in the form of interactions between ants. There is a queen in the colony, but she does not direct the activities of the individual ants. The queen is merely an ant factory. She is the sole repository for producing more of the colony. The queen, rather than being a ruler of ants, is a kept woman. She is enslaved by the colony for a single, vital purpose. In the same way, DNA is a kept molecule, not a ruler. It is a repository for making more protein. All of the actual work is done by protein. It is protein that makes DNA, and it is protein that makes RNA, and it is even protein that makes protein, in a sense. The information of a protein colony exists largely in the form of interactions between molecules as much as within the molecules themselves. Most of the interactions are

between protein and itself. DNA is clearly not the center of the molecular information universe. Perhaps it is helpful then to think of protein as more like the sun, but protein and DNA must always be seen relative to each other, and together they move through the molecular information universe.

The information to make protein lives in protein, but it is stored in DNA. Of course new protein must be made all the time, both in the sense of producing existing forms and in the sense of finding new forms. This molecular building system is an integrated production, search and storage system. It is only by considering it on the right scale and with the proper perspective that we will finally realize this. The effort begins with a simple correction to the onerous central dogma, a key part of a defective model that had us looking and thinking in all the wrong ways.



Self-reference is the real key to molecular information, and the majority of it is centered in the protein sphere. Once the proper time perspective is appreciated, we can begin to make proper analyses of molecular information and its "flow." Consider the information in a single DNA chain that will become a protein. We can track this information through all three orders of molecular information. On the first-order, the information is translated into many mRNA molecules, and each of them is translated into many protein molecules. Note that the information in DNA is tightly compressed in many ways. The translation process unpacks and expands the information in various and complex ways. First, the structures contained in the DNA chain are exquisitely precise. They are expanded in space and time by a set of more robust, larger structures called tRNA. Once the DNA information is imparted on amino acid chains, it immediately begins its interaction with the information contained in all of the other molecular structures in the environment. It is only because all of the structures are integrated parts of the exact same system that there is any consistency to the interactions whatsoever. The question of how random, complex assemblages of atoms can become predictable becomes many questions about molecular information, what it is, where it is, how it behaves. The answers will be complex, not simple. We are decades away from having meaningful answers, obviously.

Eventually, all of the molecular information within the environment reaches a point in time when it is necessary to make more DNA molecules. The aforementioned sequence in question must be "reproduced." However, we can now in this way say that one sequence of DNA leads to many other molecules, which consist of RNA, protein and more DNA. The information in DNA exists in a compressed format and it is decompressed in many complex ways. The second-order of time-dependent molecular information represents but a few of them. First, the DNA sequence is replicated, obviously. Second, the DNA sequence is combined with other sequences. Third, the

DNA sequence is transformed. In other words, the information in a single sequence of DNA expands on the first-order, and it expands on the second-order as well. Perhaps the most ingenious expansion, however, can be glimpsed on the third-order of molecular information. It is when we consider our protein system as a whole that we recognize the ultimate compression within a single sequence of DNA. The symmetry of DNA base pairs and codon triplets is reflected in the map between nucleotides and amino acids. What could be more precious? What this means is that each DNA sequence is seen by the system as many unique DNA sequences! Every sequence has a complement, an inverse, frame shifts and point mutations. The symmetry of DNA and codons is leveraged by the third-order organization of the entire protein system. The protein search space directs the protein search. Where, in this universe, is there more self-reference?

It is perhaps now slightly easier to appreciate the awesome consistency seen in the various parts of the system across trillions of trillions of cells that might otherwise have spun off in all manner of evolutionary paths billions of years ago. In other words, it is not a simple, one-dimensional, arbitrary, inefficient, linear information system that is universal because it was magically and illogically frozen out of the evolutionary game by a mystical "functional imperative." Instead, the genetic code is a breathtakingly efficient, ingenious, complex, logical, multi-dimensional, non-linear information management system that has been hotly evolving since the first days of earth. It actually displays many and fascinating slight variations on many and complex levels, but the system itself is remarkably consistent across all life. The consistency we see in the genetic code reflects the consistency in the notion that life has a shared goal, and this goal can only be achieved by life collectively. Life is generating molecular information of increasing complexity and at exponential rates through time. This is what life does. Life accumulates information.

Wrapping up a few of those pesky definitions

I hope that the intrepid reader who has made it this far realizes that we have some extremely serious problems with the consensus ideas about the genetic code today, and that these problems manifest in the definitions that we casually use for key words used in the common narrative of this "molecular information" system. You cannot be almost pregnant, and you cannot be half-dead (although in the emergency room I encounter this phenomenon more than I care to). The point is, the genetic code is either one-dimensional or it is not. Molecular information is either sequence or it is not. A protein is either merely a sequence of amino acids or it is not. The central dogma is either valid or it is not. Note, however, that these are all just different ways or articulating the same fundamentally bad idea. Everything we like to say about proteins, the genetic code and molecular information is just mutually supportive hogwash. If one falls they all fall. We need pull only one thread and the entire thin tapestry unravels completely.

It should be painfully obvious from what I have written here that it is well past time to pull the thread and let the destructiveness of our ill-conceived false belief system finally begin to dissolve. For goodness sake, there are thousand of thick, colorful threads everywhere we look, and a new, vibrant, warm and exciting cloth of insight and

discovery will be woven from them in the near future… I hope. And it can all start with a few, simple, "good" definitions. Here are just a potential handful:

Molecular information – molecular structures plus time-dependent structural interactions between atoms and molecules.

Amino acids – A set of molecules based on a central carbon tetrahedron, substituted at two vertices by a carboxyl moietie and an amino moietie and capable of forming canonical amino acid chains.

Canonical amino acid chains – chains of amino acids that are connected by carboxyl-to-amino linkages.

Protein – A semi-permanent structure from any assemblage of molecules consisting of one or more canonical amino acid chains.

Nucleic acids – A set of molecules based on carbon and carbon-nitrogen rings that are generally capable of forming chains and pairing with complementary rings in structures with icosahedral symmetries.

Nutein – A semi-permanent structure from any assemblage of molecules consisting of one or more nucleic acid chains.

Pronutein - A semi-permanent structure from any assemblage of molecules consisting of one or more nucleic acids and one or more amino acids.

The Genetic Code – An anachronistic misnomer that was erroneously defined as the function between codons and amino acids, but can now properly be defined as the molecular language between nuteins and proteins.