

Four Laws of Biology Used to Locate the Genetic Code.

By Mark White, MD
Copyright Rafiki, Inc. 2008

Introduction

The concept of a genetic code of protein translation was invented in the mid-twentieth century. It is now a central model of molecular biology holding that one dimension of molecular information is stored in DNA and translated into protein. This model is embodied in the standard table of codons that simply map sixty-one codons to twenty amino acids, and the remaining three codons are mapped to stop signals. Central to this model are three mutually supporting axioms:

1. There is a single “target” structure for all thermodynamic protein folding.
2. The genetic code is a simple co-linear process, transmitting a single dimension of molecular information from DNA to protein.
3. The linear molecular information “flows” in only one direction, according to the central dogma of molecular biology.

All three of these axioms are patently absurd, empirically false, and merely tautologically structured to sustain the model. Unfortunately, this simple model structures our current web of beliefs in molecular biology, yet it has utterly failed. Therefore, the model and its resulting web of beliefs must be explicitly rejected and replaced by something meaningful. However, to fully understand this medieval backward situation of science we find in molecular biology today – ironically one of the hottest fields of modern high technology - we must examine the path that brought us here.

Science is always complex, never simple; nonetheless, we can often detect a simple pattern underpinning its advance. Scientific beliefs generally start with observations, which lead to inductive conclusions, which further lead to deductions, then more observations, and eventually the process manifests as a coherent web of mutually supportive beliefs that guides our thinking and delineates the performance of still more science. In order for this system to be effective, however, our inductions and beliefs must somehow be falsifiable. Therefore, science must be ever-vigilant to false beliefs, yet the current situation regarding the universally accepted concept of a simple, linear genetic code is clearly not following these basic rules. It is a monstrous outlier in the history of science. Or is it? Regardless, we can trace the problem back to the invention of the genetic code.

We know that proteins are comprised of twenty standard building blocks called amino acids. We know that the bonds between these building blocks are variable with respect to their general conformation. We know that there are on average about three hundred amino acids in a typical protein. We know that, given the nature of variability of conformations between amino acids that a large measure of variability in the overall conformation of a typical protein can be expected. All of this was known before the current model of the genetic code was even proposed, but the current model now denies

the simple concepts behind our logically unassailable foundational knowledge. Here is the sequence of events that led to this indefensible situation:

Observation: An atypical protein (i.e. a relatively small one) was shown to possess - and seemingly return to - a consistent conformation under a range of atypical circumstances.

Induction: Proteins can be understood to have only one stable conformation (native state) and all other conformations can be considered inherently unstable (random coil). Therefore, all proteins in the random coil state should be seen to be thermodynamically moving toward their inevitable native state (protein folding).

Deduction: Amino acid sequence determines protein structure, so molecular sequence defines molecular information.

Deduction: The genetic code is one-dimensional, and the one dimension of this code can be perfectly embodied in the standard codon table; therefore the code itself is simple, can be located within an ordinary spreadsheet, and can be completely understood in a matter of minutes.

Deduction: Molecular information can only flow from nucleotides to proteins.

Deduction: Since most species use the same genetic code spreadsheet - more or less - most species use the exact same genetic code.

Deduction: Since the genetic code is so simple and mostly universal, and since it is a required trait for all living cells, the genetic code has barely evolved since its appearance on earth. Furthermore, it must now be considered as an obvious and prominent inherited living trait of the last common, single-cellular ancestor of all life on earth.

Deduction: Since the genetic code is prohibited from evolving by its “functional imperative,” and since circumstances on earth have changed so dramatically since its miraculous and sudden first appearance on earth, many of the features of the genetic code can be seen as being more or less arbitrary today. They are simple “frozen accidents” within an inefficient legacy system of protein translation.

Observation: Chaperones are required for proper folding of many proteins.

Observation: “Mishappen” proteins - called prions - are demonstrated to cause disease.

Observation: “Silent” mutations change the shape of folded proteins.

It should be quite obvious from this account that there is a serious problem with our fundamental model of the genetic code, the simple model lying at the heart of molecular biology and founding the core of our web of beliefs in this field. Note that it all began when a relatively weak observation was made, one that was entirely

counterintuitive to the foundation of our existing knowledge. This observation was then used as the basis of an incredibly strong inductive conclusion. This conclusion was then used as the premise for many highly related deductive conclusions that merely added to and actually built a semi-coherent yet impenetrable web of belief. An entire model and a specialized language to communicate this model have resulted from a single false induction. This is the state of our current model, a central concept in all of biology. Today it stands on no logical or empiric foundation whatsoever. In fact, the original “Law” of linearity has been falsified so many times and in so many ways that it is now hard to imagine any more ways that it can be falsified. There are no observations that support the radical inductive reasoning that gave it purchase in the first place, that forced us to ignore common knowledge, ignore intuition and ignore basic logic. The whole model is beginning to resemble a gigantic scientific hoax, so what is not quite obvious is why this situation is allowed to continue so long and so completely unchallenged.

The answer to this enigma is that these things are now central to our beliefs and without them our web of beliefs will perhaps collapse entirely. Another explanation is that the language is so finely programmed to these beliefs that the language itself prevents us from abandoning the beliefs. So, in order to replace the web of beliefs, we must first replace the “Laws” that found it, and then propose a new system of communication that can support a new web of beliefs that will inevitably build up around them.

Science concerns itself with four basic questions: What? How? Why? And Where? When it comes to answering questions about the genetic code, the classic model is marginally good at completely answering three of them tautologically: What? How? Why? Curiously, the fourth question never gets asked, and that surely is going to be the ultimate downfall of the classic model:

Where is the genetic code?

Four Laws of Biology

To begin answering important basic questions about the genetic code, we must first have “laws” that guide our thinking. We will start replacing our failed model of the genetic code with a new one by first proposing four laws of biology. Each law will be described in some detail, and then we use them to roughly shape a new model of the genetic code. The laws and the new model can then serve as a foundation of language that might support and communicate them.

- Law #1: The law of comparison.
- Law #2: The law of consistency.
- Law #3: The law of complexity.
- Law #4: The law of compression.

These four laws are mutually supportive; one might even say that each builds upon the previous. All of these laws have perhaps been stated elsewhere in various forms, but we will try to make them into a coherent whole here specifically for application to the genetic code.

Law #1: The law of comparison.

Biology is organized by evolution via natural selection. These are the things that organize the biosphere and therefore organize our thoughts, nomenclature and our system of communicating concepts within biology. Natural selection is a system of pure comparison. First, note that natural selection is crudely described by the phrase “survival of the fittest.” But what determines the state of being fittest? The answer: compare all things to all other things and then arrive at a determination of each individual state of *relative* fitness. In other words, natural selection is an inherently comparative algorithm, or if you prefer, it is an ingenious sorting algorithm of uncanny effectiveness.

Second, we might speak of various languages of life or languages in life, but we must realize that all languages are based on metaphor. Metaphors are mechanisms that derive meaning by comparing one thing to another. Therefore, all languages are also inherently comparative. Indeed, life can be thought of as a language in and of itself.

Third, symmetry is an essential concept toward our understanding of the universe at all conceptual levels. It is no surprise then that we should find symmetry everywhere within living systems. But what is symmetry? Once again, symmetry is pure comparison. Symmetry is the state of things when one thing is transformed into another, yet the two things somehow remain the same. Symmetry is change without change. Symmetry is reversibility, from change to unchanged. Symmetry is inherently a comparison of two things, one thing before change and the other after.

Fourth, information theory now roams free at the heart of biology. What is information? Once again, information is comparison. It can be stated in many forms, sender versus receiver, intended message versus interpreted message, choice versus choices, it all comes out the same; it is one thing compared to another. All calculations of information are calculations of probability, or ratios of a choice made versus the probability of any choice made.

The law of comparison holds that all concepts central to biology ultimately come down to various forms of comparison.

Law #2: The law of consistency.

This Law builds upon and expands the first Law. We have acknowledged the central importance of natural selection in biology, and we have noted that natural selection is a sorting algorithm. But on what criterion does it sort, or what is it seeking? The answer is that life is seeking optimums. It seems reasonable to think that optimums imply consistency. In other words, there is always only one best way. This principle has been stated and restated in many ways. Perhaps the first, deepest and most eloquent statement of this idea was made by D’Arcy Thompson in his seminal work, *On Growth and Form*. Thompson noted that nature is constrained by physics, and so some physical forms are required by life – optimums - so by the law of consistency these forms can be expected to consistently show up over and over again. However, we should now also note still more abstractly that the same thing must be said at the molecular level, and it should also be said about the basic numbers that show up within life’s many processes.

There is only one set of natural numbers, there is only one set of logical relationships between these numbers, and so those numbers and their fundamental relationships can be expected to repeat themselves throughout life. This is perhaps most prominent in the multiple occurrences of the Golden Ratio, but we can also expect consistency to arise at still more primitive levels within life's processes; from no other logical obligation than optimization.

The law of consistency states that life seeks consistency in its numbers, forms, and relationships as a way to achieve optimization.

Law #3: The law of complexity.

Perhaps the best way to state the law of complexity is to state its negative: life is never simple. However, the concept is surely more profound than this. Life uses complexity as a means to advance life. Complexity is both a means to an end and an end in itself. Perhaps we can comprehend this principle by noting that life builds tomorrow with the tools it finished building today. Therefore, the things built tomorrow will logically become more complex. One might say that life's goal is complexity and life employs complexity to achieve its goal. From this perspective, we can note that life's goal today is to build better tools for building more complexity tomorrow. Life's favorite tools then would be the ones that can provide the most future complexity.

This concept builds upon Laws #1 and #2. By comparing the utility of potential tools, and finding the most basic (i.e. the most consistent) tools for building, life is able to combine things into more efficient systems for future building.

Another way to restate this law is to say that life cares not for reproduction; life cares only for production. Life seeks non-linearity not linearity. Reproduction is simple and linear whereas production is non-linearly complex. Life uses reproduction merely as a tool to achieve production. Life is seeking novelty, not repetition. Life is a shuffle sort, where existing features are combined, recombined, compared and sorted. This is the type of system that thrives on complexity, not simplicity, and the kind of system where sub-systems emerge as a hierarchical organization of lower sub-systems. Life is a fractal pattern generated by a fractal algorithm: natural selection.

The law of complexity states that true complexity underlies all apparent biologic simplicity.

Law #4: The law of compression.

This law is the culmination of the first three. The law of compression might go by many names, but to be perfectly honest, I picked this particular name to maintain a pure alliteration in the names of the four laws. Perhaps we can call it Schrödinger's law, or simply the law of negentropy. Erwin Schrödinger was prescient and eloquent enough to suggest this law in the early to mid-twentieth century, a time before molecular biology jumped the conceptual rails. Schrödinger noted that life must have a way to combat all the entropy that it inevitably must gain while alive. He therefore coined the term negentropy to describe this.

Thermodynamic entropy and information entropy are two statements of the same concept. Information, we have already noted, is a comparison, or perhaps an actualized probability. Thermodynamic entropy, on the other hand, predicts that a closed system compared at two instants in time will always demonstrate a decrease in organization through time. In other words, the inherent organization of physical systems naturally decreases with time. The law of compression states that living systems follow a diametrically opposite path. Evolution is the organization of information through time. Life is evolution. Life is increasing the organization of life through time. In other words, the amount of information in life is increasing with time, and this trend is accelerating.

Just as we see chemical catalysis as a hallmark of life, we can see information catalysis as a hallmark of living information systems. Chemical catalysis moves forward the time of expectation for a reaction. Negentropy moves forward the time of expectation - increases probability - of an event. A single hemoglobin molecule, for instance, is thermodynamically impossible in the absence of a system that contains the information needed to create it. The existence of information leads to the formation of more information and thereby increases the overall information of the system. It is a positive feed-forward loop.

Another way to see this concept is by realizing that life seeks compressible information. It uses this form of information for storage, and it decompresses this information for usage, just like a digital camera. To understand this, one must consider the storage information simultaneously with the information contained in the algorithms that pack and unpack it. Life compresses the data and the algorithms. Another example of this is a human zygote. A zygote can be thought of as the compressed information for generating a more complex organism, but where is the information that is needed to turn this cell into the final organism? The answer is that it is not in the cell, but it is incrementally decompressed from the cell. In other words, if it is a 100-step algorithm to go from cell to organism, only step one is contained in the cell, and step 2 merely emerges from execution of step 1. The entire data and algorithms do not exist in the first cell but can be derived from it. Therefore, the "data" for this procedure represents but a fraction of the total information contained in the total procedure, the rest of the information is compressed within the algorithm that uses the data. We must therefore understand not only the evolution of the data but also the evolution of the algorithm in parallel.

Another good example is the translation of DNA into protein. The information for making a protein is compressed within the DNA that leads to protein and within the algorithm that operates on the data. To fully understand the information contained in the final protein we must understand the algorithm that compresses the data into DNA as well as the algorithm that decompresses it back into protein. They are symmetrical processes, and neither can be properly understood without knowledge of the other.

The law of compression states that information within living systems expands through time, at all levels, and at an accelerating rate.

Application of laws to the genetic code

The irony is not lost on me that displacing the codon table as the embodiment of the genetic code begins with an examination of a new codon table. However, nothing communicates more effectively the deficiencies in the old model than an examination of the applications of the four laws of biology to this particular codon table.

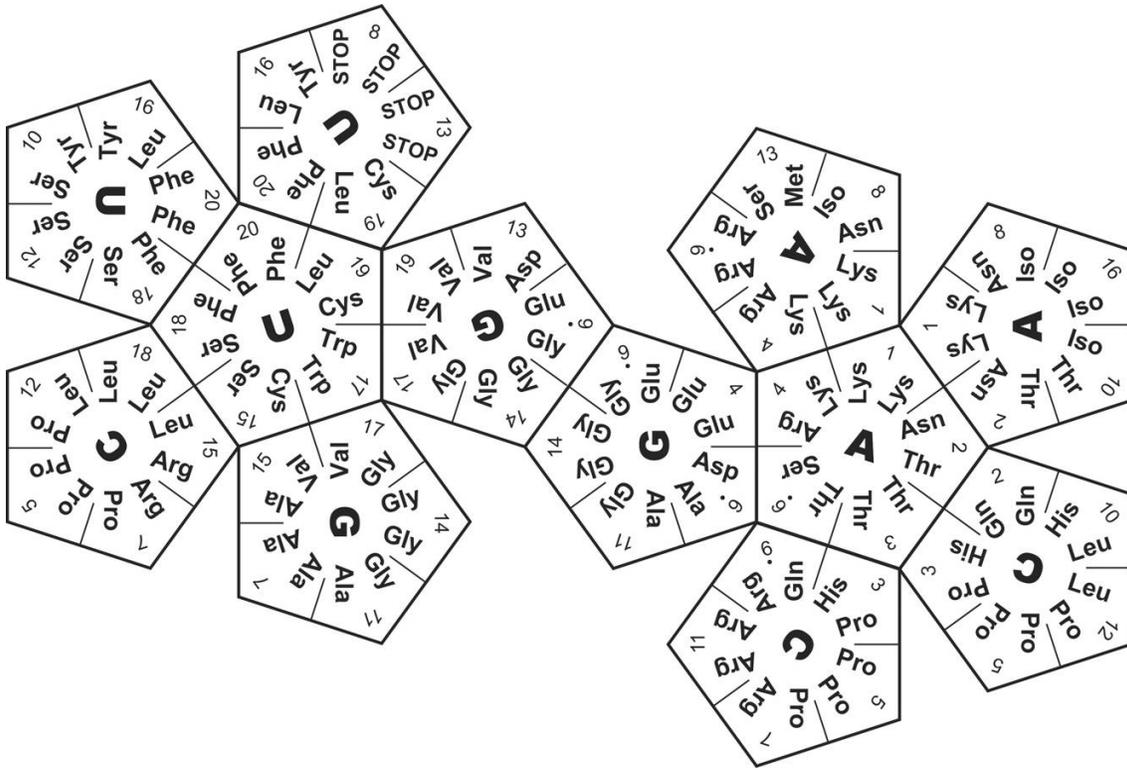


Figure 1. The Rafiki G-Ball

This codon table is explained in detail elsewhere; however, we can use it here to make a few simple points. First, this arrangement vividly demonstrates the Law of comparison. This data treatment shows that all nucleotides derive their “meaning” by comparison to other nucleotides, all codons by comparison to other codons, all amino acids by comparison to other amino acids, and finally, codons are compared to amino acids. However, even this table is a crude conceptualization, and the meta-meaning of any table, so to speak, is that whole collections of codons only have meaning when compared to whole collections of other molecules, including amino acids. Also, since symmetry is comparison, it’s hard to beat this perfectly symmetric treatment as a means of comparison.

The law of consistency could not be more clear than in the existence of a codon table. Life repeatedly uses consistent numbers, forms, and actual molecules; if not, this presentation would not be possible. The fact that we can make any presentation of this information at all should give us pause, but this particular presentation begins to make

sense of our ability to “know” this data. It is a direct product of the Law of consistency. Life is optimizing the system at this level and has been doing so since the beginning of life. We should therefore expect quite a bit of consistency at this level. It is not a measure of life’s inability to change but a measure of the futility of change at this level.

There are consistent forms: dodecahedrons and tetrahedrons, that have consistent relationships, that logically lead to consistent numbers, that lead to consistent sets of molecules that optimize those foundational consistencies. In this light, not only can we say that two base-paired sets of nucleotides grouped in threes provides the most efficient symmetry within a system of twenty amino acids, but we can say that it is the consistency of these foundational relationships that selected for these particular sets of molecules, and not vice versa. In other words, we should right the ship of cause and effect with respect to any codon table: The table chooses the molecules, the molecules do not choose the table.

The law of complexity is also apparent. Rather than a simple spreadsheet viewed at one level of meaning, revealing the infrastructure of a simple, linear code, the codon table reflects a complex organization of a complex hierarchy of molecular relationships that comprises a complex code of molecular behavior. The genetic code, rather than being absurdly simple, is perhaps the most complex thing on earth. It is a case where superficial simplicity merely masks a true, underlying foundation of complexity. It is only through application and re-application of principles of complexity can the apparent simplicity emerge in the form of extreme consistency. In other words, consistency of protein form is not a cause leading to this system but a result of the system. Highly limited molecular sets are not given but found. Every nucleotide, amino acid and robust protein and combination thereof must be found and selected for or against. Consistent protein forms are a selected feature of the system, not an a priori feature of the system. The apparent simplicity of current proteins is the end result of a long process of complexity, not a simplistic “law” given to a system at the time of inception.

The law of compression is demonstrated in a literal fashion by this particular treatment of the data. This is the most compressed format possible. There are only twelve nucleotide symbols in use, and that is the minimum required. This demonstrates that the data is not only compressed but compressible. Of all the possible arrangements of these two sets of molecules a vanishing few of them will be consistent with and compatible with this particular format. We can either conclude that it is coincidence, or that it is cause and effect. It seems obvious that the law of compression dictates that we assume the latter and not the former.

By tightly compressing the translation algorithm of proteins, life is able to store protein information in the most compressed possible format. It then uses compressed algorithms to decompress stored information most efficiently whenever protein is needed. Additionally, but much less obvious here, this scheme is also part of the optimum scheme for finding new proteins as a need for new proteins arises. In other words, the genetic code is not only an elegant operating system, it is an integrated and powerful search engine. It is hard to imagine a more perfect compression-decompression algorithm on earth than the genetic code. It is not a living example of simple, universal, arbitrary frozen accident. It is a perfect example of a complex, variable, hotly evolving and near perfect information handling system. How could it be anything else?

Having now identified the basic problem with the old model and laid the groundwork for a new model to replace it, we shall return to the four questions science must answer: What? How? Why? Where?

What is the genetic code?

The genetic code is a globally symmetric and sophisticated natural comparison between nucleotides and proteins. It is complex, non-linear, variable from species to species, yet remarkably consistent due to a highly evolved optimization of its core processes.

How does the genetic code work?

The genetic code stores information in highly compressed formats within DNA. The code decompresses this information into complex protein populations through a complex sequence of molecular events that include DNA, RNA and protein, operating within a backdrop of time and space.

Why does the genetic code appear the way it does?

The genetic code embodies the central processes of life on this planet. Without it there is no life on this planet. It is a molecular information handling system that uses comparison, consistency and complexity as its primary tools. The goal of the system is to expand information within the system at every level of time and space, and it is able to achieve this goal at an accelerating pace. The genetic code appears the way it does because it is the most critical process and the most highly evolved process within all of life. It is the living embodiment of all the laws of biology.

Where is the genetic code?

This is by far the most difficult question to answer. We will need additional conceptual tools to help answer it. Before we can compare two things we must have two things to compare. We have defined the genetic code here as a comparison between two kinds of molecules: nucleotides and proteins. What are molecules? There are no simple answers at this level, so we need a new scheme for giving meaningful answers. Whenever we speak of any molecule we must at least answer three questions about the molecule:

What is the level of complexity we seek?

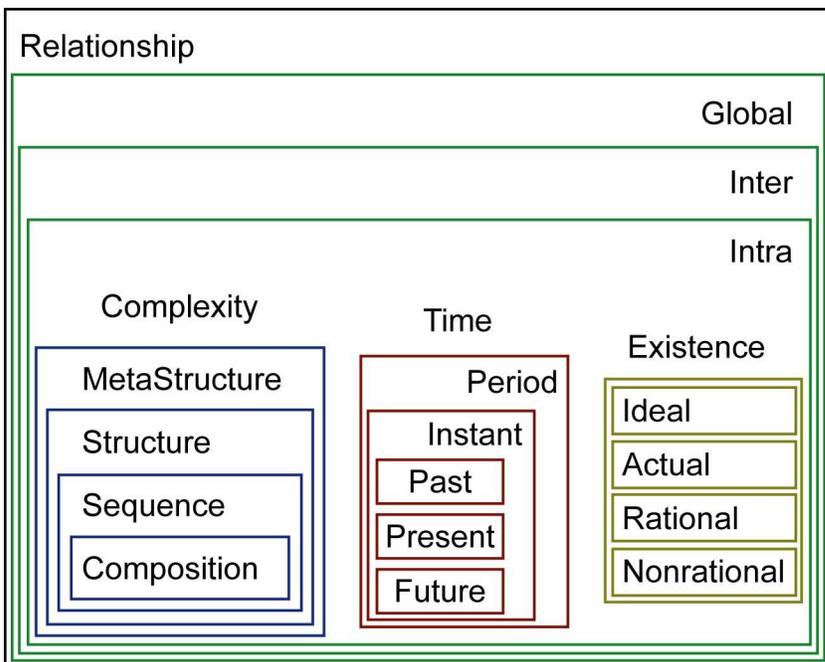
What is the time scale on which we are speaking?

What is the nature of the “existence” of the molecules to which we refer?

Once we identify each of these things for each molecule, we can then identify the nature of the molecular comparisons. Only then can we hope to understand the nature of a comparison between two molecules and the information contained therein. To help

visualize this process we have prepared a matrix of descriptors that must be used for any molecule.

Information Matrix



What is true for a molecule is also true for comparisons between molecules, and the genetic code is a comparison between molecules. We can make comparisons on many levels, so we can conceptually speak of many different genetic codes. The “ideal genetic code” is a global comparison between “all possible” proteins and “all possible” nucleotides. This code can be located within the fabric of the universe. Surely it represents the pure geometry of space and time. The rational genetic code is represented by the inter-relationship of all genetic codes past, present and future that must execute within all cells on earth. An actual genetic code can be found within the molecules of any cell. Different sets of molecules lead to different dialects of the code. So, we can locate different forms of the genetic code within individual cells, all cells, and all possible cells, depending on which genetic code we mean to discuss.