Language Evolution in Humans and Ancient Microbes: What can human language acquisition tell us about the origin of genetic information?

Stephen J. Freeland and Melissa Ilardo

University of Hawaii, NASA Astrobiology Institute

freeland@ifa.hawaii.edu

Abstract

This paper seeks to encourage dialog on a question with a deceptively simple surface. When we find linguistics and genetics using the same vocabulary to describe their central phenomena, is this because the phenomena are meaningfully similar? Are we encountering a superficial analogy, enjoying the benefits of a good metaphor or recognizing some deeper principles of information organization and transfer? We approach this broad topic by focusing attention on the ancient evolutionary events that created a system of genetic coding, soon after the origin of life on our planet. Specifically we examine a progression of three topics: whether genetic code-words are arbitrary signifiers for the objects they encode (amino acids); how evolutionary biologists have deduced clues about the evolution of genetic coding by studying the complex end-product; and the current scientific paradigm for the origin of genetic information. Our suggested points of connection suggest encouraging insights from each side (linguistics to evolutionary biochemistry and vice versa), though our primary aim is to ask for further help exploring how linguistics can reshape thinking within evolutionary biology.

Keywords: evolution; genetic code; amino acid; alphabet; translation; metaphor

Background

"Whether we realize it or not, translation is the hidden lens through which almost all of human knowledge is processed. Pick any subject – religion, philosophy, science literature, history – and somewhere at its heart is a foundational work of translation".

With this claim, Filkins (2012) summarizes an argument that a proper understanding of translation underpins academic scholarship. The argument's literal point is easily illustrated with reference to a few well-known examples. The rational philosophies of ancient Greece passed through the Golden Age of Islam en route to shape the renaissance of Western Europe, and the Hebrew text of Genesis has flowed through numerous other languages before converging to colour contemporary political debate within the USA. In each case, an original text has undergone a process of translation in order to influence cultures widely separated in space and time. As the text moves between languages, the act of translation brings an inevitable shift in semantic content simply because each language exhibits a unique association between concepts and their corresponding expression (see, for example, Armstrong 1993 for a description of the ways in which culture, particularly religion, influenced translation of Greek philosophy; Munger, 1999 provides an excellent introduction to the challenges posed by biblical translation). Indeed classical

linguistic theory teaches that even within a single language, an imperfect mapping of words to concepts forces us to undertake moment-to-moment acts of translation as we struggle to communicate our thoughts (e.g. Saussere, 1983). It is the last of these insights that forms a start-point as we now seek to travel beyond the literal truths of Filkins' quote in order to explore whether something similar holds true for the seemingly distant research domain of ancient biochemistry.

The idea of translation as a subtle modifier of meaning is the foundation from which we build, writing as evolutionary biologists who research events that occurred close to life's origins. Specifically, we are interested in the origin of genetic coding – also known as gene translation. We offer an account of what is known (and what remains unknown) about the evolution of genetic coding, with a focus on three specific features: (i) the relationship between genetic code "words" (codons) and their "meanings"; (ii) the extent to which key features of genetic coding have been optimized by natural selection; and (iii) explanations for the origin of such a system. Our intention is to solicit input from linguistics for new and fruitful ways to understand the language-like properties of the biochemical system invented by early biological evolution. With that in mind, we aim to introduce the relevant molecular biology in a manner that is accessible to an audience of non-experts, and ask you to keep in mind that it is precisely our inexperience in the field of linguistics that leads us to seek input.

We finish this introduction by noting that we consider ourselves astrobiologists – that is, participants in a multi-disciplinary scientific community that seeks to place its specialized research findings into a broader context of addressing questions of life's origin(s), distribution and future in the universe(see: Des Marais, Nuth, Allamandola, Boss, Farmer, Hoehler, Jakosky, Meadows, Pohorille, Runnegar & Spormann 2008). Beneath our explicit research question lies a motivation to clarify the roles of chance versus predictability in the evolution of life on our planet. One might infer that the existence conscious beings capable of asking and communicating such questions is the sum product of countless lucky but arbitrary events rippling into ever more unpredictable consequences through time and space (Gould, 1989). An opposite conclusion is that our species' existence is a predictable outcome, given the basic chemistry and interactions inherent in the universe (Morris, 2003). The truth seems likely to lie somewhere between these two extremes, and the scientific challenge is to find where, exactly. More accurately, we perceive the challenge is to find creative ways in which we can begin to refine estimates (see, for example, Freeland 2007). This is the theme to which we return in our concluding comments.

An introduction to biochemical language

Contemporary science uses simple linguistic vocabulary to describe the interface between biology and chemistry. All living systems "read" and "write" genetic information using appropriate chemical "alphabets". Above all, DNA genes are "transcribed" into RNA and then "translated" into protein, an entirely different chemical "language." Proteins are the metabolising entities whose functions define life as we know it, including the ability to copy genetic information into a new generation (see Figure 1).

The discoveries contributing to this "*central dogma of molecular biology*"(Crick, 1958) marked a mid-twentieth century turning point for biological science, and were rewarded with an appropriate shower of Nobel prizes (including one each for the localization of biological heredity within DNA(1969, Physiology or Medicine), the structure of DNA(1962, Physiology or Medicine), the process of genetic coding(1968, Physiology or Medicine) and that of protein folding(1972, Chemistry)). Collectively these achievements have made possible the biotechnologies of gene sequencing (Maxam & Gilbert, 1977) and genetic engineering(Hughes 2001). They have enabled scientists to reconstruct the evolutionary tree of life back to the Last Universal Common Ancestor (LUCA) of all today's living species (Williams, Fournier, Lapierre, Swithers, Green, Andam & Gogarten, 2011).

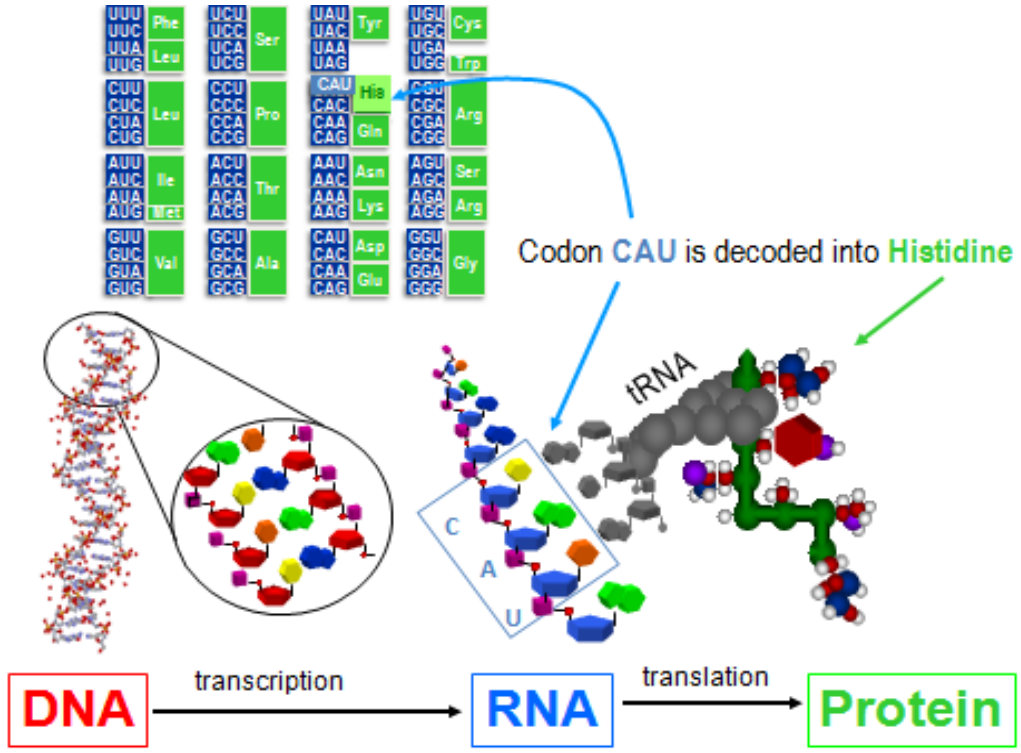

Figure 1. For at least 3 billion years, all living systems have consisted of a process by "translating" genetic information, "written" in the form of nucleotide sequences (chemically bonded into a linear chain known as nucleic acid) into corresponding sequences of amino acids (again chemically bonded together into a linear chain). A given sequence of amino acids will spontaneously fold up into a 3 dimensional shape, a protein that exhibits specific structural and functional properties, such as catalyzing a specific chemical reaction. Networks of protein molecules thus interact to form life – including the processes by which nucleic acid information is copied into a new generation through reproduction.

The genetic code nothing more and nothing less than the set of rules defining exactly what each possible genetic code-word (codon) is translated into. All codons are exactly 3 nucleotides in length, thus the genetic code is usually depicted as a table in which each possible codon (3 letters of RNA – a minor chemical derivative of the more famous DNA).

The genetic code is put into effect by a class of molecules known as tRNA's. Each tRNA is an adaptor molecule: at one end of the tRNA, an "anti-codon" recognizes a specific codon through chemical interactions; at the other end, the tRNA carries a specific amino acid. In this way, a suite of tRNA molecules enable translation of a specific gene sequence into a correspondingly specific sequence of amino acids.

Quite beyond this impact on human knowledge, the legacy of the central dogma awarded a far greater importance to the microbial lineage(s) that first evolved genetic coding, the organisms at the forefront of this historical transition. Current data indicate that the translation of genetic information into proteins, with all its language-like qualities, appeared more than 3 billion years ago (that is, around three quarters of our planet's history) (Fox, 2010). The result was a system of such evolutionary plasticity that biologists of the twenty-first century are struggling to find physical conditions that represent boundaries to life(Des Marais *et al.*, 2008). Already so-called extremophile organisms are known to have adapted to life at the deepest depths of the ocean and the highest reaches of the atmosphere (Satyanarayana, Raghukumar and Shivaji 2005). They divide and multiply in boiling water, deep-frozen ice, lakes of acid and streams of alkali. They survive extra-terrestrial trips on our space-craft and even grow on spent nuclear reactor fuel. It is not difficult to understand why luminaries of evolutionary biology list the advent of genetic coding as one of the major evolutionary transitions for life (Maynard Smith and Szathmary, 1995), ranked alongside life's origin and the emergence of language for its contribution to our current existence. Understandably

then, research into the origins of genetic coding is a well-populated field with a correspondingly large scientific literature. Yet despite the linguistic terminology used by biologists, the language-like properties of genetic coding rarely merit serious consideration – or even an overt acknowledgment that something here deserves exploration. Is the trans-disciplinary vocabulary meaningful or misleading? Are we creating potential confusion with a careless analogy, or have we identified a useful metaphor that allows useful transport of insights and ideas across a large interdisciplinary divide?

Genetic coding and arbitrary signifiers

The key to any form of translation, including that found at the heart of biochemistry, is a code. A code is a system of rule(s) for converting one piece of information (for example, a letter, word, phrase, or gesture) into another representation (see for example Oxford English Dictionary, online edn). In other words, codes are interfaces that allow semantic information (useful meaning) to flow between languages. For example, when we talk about the Morse Code, we refer to the system of rules that describes how to transfer meaning between a simple alphabet of dots-and-dashes into the richer language of letters and symbols with which this paper is written. Likewise, the genetic code is the interface between a simple alphabet of 4 nucleotides and the richer alphabet of 20 amino acids (Figure 1). Just as Morse code assigns code-words constructed from the simple alphabet to a single letter of the more complex alphabet (for example *dot-dot-dot* is the code-word assigned to the letter *S*), genetic coding assigns nucleotide triplets, known as codons, to a single amino acid (such as *A-A-A* used to encode the amino acid *Lysine*).

In this sense, popularized science feeds confusion when it uses the phrase "genetic code" to refer to the genetic information stored within DNA (see, for example, "*Cracking the Genetic Code of Brown Tides*," Rudolf, 2011)). DNA is *not* a genetic code: rather, the genetic code is what imbues DNA with useful information. The information content of DNA (or any other language) is entirely defined by the code at work. Whether we consider a gene written in DNA, a gesture passed between a dog and its owner or your interpretation of this written sentence, changing the code changes the information. At the risk of labouring this crucial point, consider the symbols:

IOI

If we are decoding Arabic numerals into the concept of a number, this could represent “one hundred and one” (under a decimal notation) or “nine” (assuming binary notation). Removing the assumption that decoding should result in a number, text-processing software using the ASCII code would perceive a binary number signifying “horizontal tab”. Dropping any assumption that numbers are involved, the symbols might represent a text message for “laugh out loud” or even the sigil of the comic book superhero Green Lantern. Semantic information is defined by the code.

From this standpoint, a fundamental question to ask of any code is the extent to which it is arbitrary. In other words, are the rules that connect two different representations of information constrained to reflect anything inherent to one or both of those representations, or to the process of translation itself? Drawing from linguistics, we suggest a metaphor of the connections between words and their meanings. At one extreme, words can be onomatopoeia: here, a non-arbitrary connection is implied as the word emulates its meaning directly. At the other extreme, words can be entirely arbitrary signifiers of the meanings they signify.

In contrast to the history of structuralism in linguistics (which seems to have *assumed* that arbitrary signifiers form the basis of language), early researchers of the genetic code *assumed* that a type of molecular onomatopoeia must be at work. Somehow, sequences of nucleotides evolved to encode amino acids. In the absence of intelligent design, how could arbitrary interactions arise? The clear alternative was to imagine direct interactions between the types of molecule involved – nucleotides and amino acids. For example, the famous double helix of DNA is a molecule with a distinct, physical structure that presents a surface rich with texture and associated chemical properties (such as positively and negatively

charged atoms). Amino acids are likewise precise molecular structures with their own distinct shapes and chemical properties. An intuitive guess was that amino acids bound to DNA through direct, complementary chemical affinities (Gamow, 1954). Great scientific minds of the mid twentieth century devised ingenious models to describe such affinities in ways that would form genetic coding as a direct by-product of physics and chemistry (reviewed with clarity and wit by Hayes, 1998).

The problem with all such onomatopoeia-like theories of direct-templating was the subsequent discovery of a type of molecule known as transfer RNA (usually abbreviated to tRNA). All organisms contain a set of slightly different tRNA molecules that are the physical manifestation of the rules of genetic coding (i.e. the genetic code). At one end, each tRNA recognizes and binds to a specific codon (or, sometimes, a suite of similar codons); at the other end, it carries a specific amino acid. The very existence of this molecule, inserted between the genetic information and its amino acid counterpart, removes any need for direct physical complementarities between nucleotides and amino acids; relationships specified by the genetic code are as (potentially) different as the association between words and their meanings. Put another way, codons (and the nucleotides of which they are composed) are, in the modern system, symbolic signifiers for the amino acids they encode.

By the late 1950's, this much was known (Hoagland, Stephenson, Scott, Hecht & Zamecnik, 1958), and the concept of a truly symbolic code is important to much of what follows below. However, we would be remiss in finishing this section without reference to much more recent findings which have resurrected the concept of direct-templating (molecular onomatopoeia). During the late 1980's, biotechnologists invented a method known as SELEX (Tuerk & Gold, 1990). The process uses *in vitro* selection to find RNA molecules known as *aptamers* that target and bind other molecules (Cho, Lee & Ellington, 2009). When multiple aptamers are selected for their propensity to bind a specific target molecule, researchers often discover common motifs occur within them all. Since individual aptamers share little or no common ancestry, these common motifs are likely responsible for their binding affinities. Interestingly, statistical analyses reveal that aptamers selected for affinity to bind specific amino acids contain motifs that are unusually enriched with the codons assigned to those amino acids in the genetic code (Yarus, Widmann & Knight, 2009). Extending our previous metaphor, it would be as if cultural anthropologists found numerous tribes had independently derived vocabulary for a dog's bark and had all arrived on minor phonemic variations of "woof".

Biochemically, this intriguing result brings as many questions as answers. So far, only around half of the 20 amino acids have been tested with SELEX (and already one has shown no interesting associations). This incomplete data set renders any firm conclusions difficult. More generally, the lesson from early, ingenious theorizing about the code is that almost any pattern can been identified if one looks hard enough. One particularly poignant example concerns a 1966 publication which used plastic models of the relevant atomic structures to reveal an excellent fit between codons and their amino acids (Welton & Pelc, 1966). Critical re-evaluation revealed, however, that the researchers "had built all their polynucleotide sequences backwards. Their AAG was, in fact, GAA" (Crick, 1967) which encodes an entirely different amino acid of entirely different shape and properties from the amino acid for which an excellent fit had been discovered.

Most mysterious is why any putative direct-templating associations should have survived the evolutionary transition to tRNA-based, symbolic coding. The insertion of adaptor molecules would clearly disrupt any mechanism of direct interactions between amino acids and their encodings. Stretching our metaphor further, it is as if human communication began with nothing but onomatopoeic utterances – and yet in some distant future, long after all human communication evolved to indirect exchange between cell phones and computers, we found these primitive onomatopoeia operating unscathed. One interpretation is that changes to the rules of a code are impossibly disruptive once it has become established – no matter how much machinery evolves to separate signifiers from the signified. This idea seems logically attractive – but it also introduces our second point of exploration for similarities between biochemical and human

language acquisition: what do we know about the emergence of a standard genetic code, shared by organisms as diverse as humans and *E. coli* bacterium?

Universal codes and footprints of evolution

A common challenge for understanding the evolution of any biological system is to find reliable clues that lead from the world we can observe backwards into the past. Broadly speaking, evolutionary research has developed two complementary approaches to find these footprints of evolutionary history, both of which present interesting parallels within the study of language.

The first approach seeks non-randomness in the form of apparent design in order to address the question, "what is this organism adapted for?" For example, we can deduce that insects which resemble twigs have been shaped by evolution to reduce their chances of being eaten by predators. More accurately, we infer that within the ancestral population from which these organisms derive, individuals with the greatest resemblance to twigs had a statistically higher chance of surviving to reproduce. Given time and the constant trickle of new, random mutations these genetic advantages accumulated into the startlingly non-random coloration and shape that first drew biologists' attention. Of course, this much alone is mere hypothesis, but it suggests a barrage of possible tests (do variations in natural insect populations correlate camouflage with survival and reproduction? Can we artificially modify insects' camouflage in order to measure the effects on predation?) In this manner, apparently purposeful design initiates evolutionary inquiries that end up revealing the evolutionary forces or 'selective pressures' from which a modern form has derived.

A less widely appreciated alternative approach searches contemporary structures for precisely the non-random features that do not make sense as adaptations – because these are often clues to an ancestral state. Specifically, features of an organism that do not contribute to an adaptation (or, better yet, those that compromise the perfection of an adaptation) can often be understood as restrictions imposed by the ancestral states from which an organism has evolved. Thus our own species' vulnerability to back and knee injuries reflect natural selection for bipedal locomotion that improvised on an anatomy previously adapted for quadrupedal gait.

Each approach has played an important role in shaping the present state of knowledge for the evolution of genetic coding. In the first category, the earliest suggestions for adaptive, non-random features of genetic coding focused on the fact that the code seems to have a built in mechanism for reducing the impact of errors. Codons that differ only at the third nucleotide position are often assigned to the same amino acid, for example codons GGU, GGC, GGA and GGC are all assigned to the amino acid *glycine* (Figure 1). Early researchers suggested this feature could have arisen from natural selection to minimize the impact of errors during translation because a high proportion of mistakes involving a single "letter" of the triplet codon have no effect on which amino acid is specified (Sonneborn 1965, Zuckerkandl & Pauling 1965). In other words, we might imagine a primordial population in which different genetic codes were at work. If occasional errors in translation were inevitable, then successful lineages might have been those operating genetic codes which mitigated the associated negative impact. That is, error minimising codes would be like well-camouflaged insects, using their adaptive advantage to out-survive and therefore out-reproduce their competitors.

This satisfyingly simple evolutionary argument was soon silenced by another discovery of Francis Crick. Specifically, his "wobble hypothesis" noted that tRNA adaptor molecules are often unable to distinguish between similar nucleotides at the third position of a codon for reasons of physics and chemistry (Crick, 1966). Under this view, the pattern of synonymous codons is a necessary byproduct of physical limitations rather than a clue to evolutionary adaptation. Crick's insight combined with his own, earlier flawed hypotheses (Crick, Griffith & Orgel 1957) to convince him that altogether too much excitable speculation had accompanied scientific discovery of the genetic code. His highly influential voice thus

went on to assert that the code was best understood as a "frozen accident" of evolutionary history (Crick, 1968). To avoid mis-representing this claim, these are his exact words:

"This theory states that the code is universal because at the present time any change would e lethal. This is because in all organisms ... the code determines ... the amino acid sequences of so many highly evolved protein molecules that any change ... would be highly disadvantageous unless accompanied by many simultaneous mutations to correct these mistakes produced by altering the code. This accounts for the fact that the code does not change. To account for it being the same in all organisms, one must assume that all life evolved from a single organism (more accurately from a small interbreeding population). **In its extreme form, the theory implies that the allocation of codons to amino acids ... was entirely a matter of chance**" (page 370, emphasis added).

It is here that we re-connect to the evolution of human language. The Frozen Accident hypothesis and its effect in silencing evolutionary hypotheses about the origin of genetic coding seem to us oddly similar to the views expressed by another academic giant, Noam Chomsky, on universal grammar and their impact on research into the evolution of language acquisition (Chomsky, 1988). Both Crick and Chomsky concluded that because of the universality of the phenomenon (the genetic code and the structure of human language respectively), any attempt to infer evolutionary history is merely unhelpful speculation.

Genetic code research has escaped from such stifling reasoning due largely to the discovery of empirical, contradictory evidence. In fact, the first example of a naturally occurring variation in the genetic code was observed in none other than our own species. All cells of the human body contain mitochondria – organelles that house their own genetic material (and operate their own genetic codes) independent of the main cell nucleus. A decade after the frozen accident was put forwards, human mitochondria were found to use a code in which two codons are assigned to different amino acids than those described by the "universal" genetic code (Barrell, Bankier & Drouin, 1979). Since then, many other exceptions have been reported in a wide variety of species (Knight, Freeland & Landweber, 2001). This has eroded credibility of the frozen accident while reinvigorating the search for adaptive features that hint at the evolutionary forces by which genetic coding emerged. Particular success has come from resurrecting and expanding early arguments for an "error minimising" code. Indeed, sufficient evidence has accumulated for this concept that a paper of this length would be required to provide an adequate overview. Happily, such reviews already exist, both in specialized academic publications (Freeland, Wu & Keulman 2003; Koonin & Novozhilov 2009) and written for interested non-specialists (Freeland & Hurst, 2004). For present purposes it suffices to give a flavour of the argument in Figure 2.

The message is that even with the simplest quantification of non-randomness, the standard (non-universal) genetic code appears quite extraordinary compared to theoretically plausible alternatives. At first sight, finding a highly optimized pattern of codon arrangements might seem exactly the sort of evolutionary overlay that would disrupt the molecular onomatopoeia of a primordial, direct-templating system. One important study offers a resolution to the paradoxical evidence that both ideas are true by showing that the vast range of plausible genetic codes offers plenty of room for both ideas: many codes could fulfill the restrictions of onomatopoeia without displaying "error minimization" properties (Caporaso, Yarus & Knight, 2005). Even the simple "language" of genetic coding contains enough flexibility to carry more than one strong signal of its evolutionary causation.

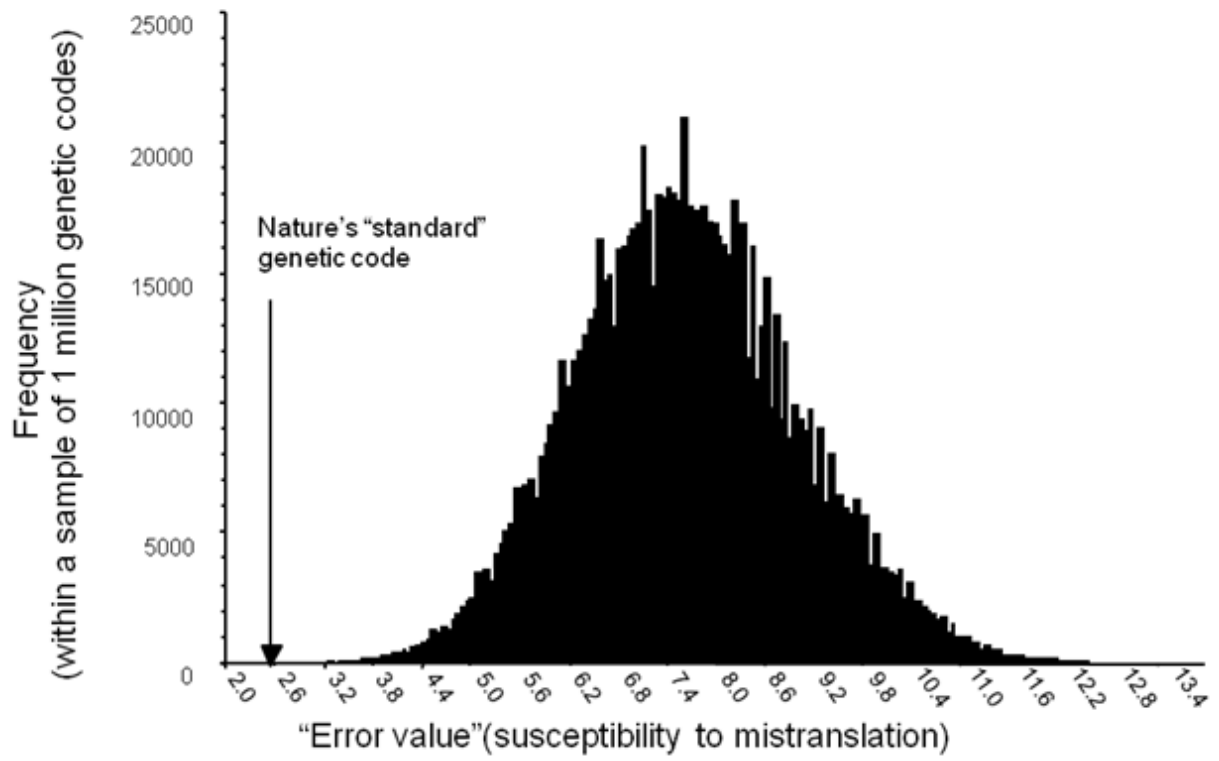

Figure 2. The standard genetic code is highly unusual by virtue of minimizing the impact of common errors that cause mistranslation. The standard genetic code is the particular set of rules that had evolved by the time of the last universal ancestor of life on Earth (and remained largely unchanged ever since).This figure illustrates a typical analysis of the standard genetic code compared to a computer-generated sample of 1 million randomized alternative codes, each considered plausible by maintaining all components of the molecular machinery used by the "real" code. Each genetic code is reduced to an "error value" that represents the code's susceptibility/ sensitivity to translation errors. The most common form of mistranslation event involves mis-reading a single nucleotide "letter" within the triplet that forms a codon. This can have a variety of effects on the fitness of an organism, according to the chemical similarity of the amino acid pair that forms the "actual" and "intended" translation of that genetic code word. A code with a large error values is one for which the average effect of common translation errors leads to a substantial change in the amino acids' chemical properties. A genetic code with a small error value is one which buffers the living system from this effect. In this particular study, within the sample of 1 million plausible alternative codes, only one was found that (slightly) improves upon the rules of the standard code in these terms.

Until recently, definitions of plausible alternative codes have focused on rearranging the pattern by which codons are assigned to 20 amino acids. This matches the empirical observations of non-standard codes that occur in nature (Knight *et al*. 2001), and increasingly robust insights into how codon reassignment takes place (Rocha, Pereira, Santos & Macedo-Ribeiro, 2011). More recently, however, a 21st and 22nd amino acid have been identified mid-way through incorporation into the genetic code, each by a different mechanism (Yuan, O'Donoghue, Ambrogelly, Gundllapalli, Sherrer, Palioura, Simonović & Söll, 2010). This in turn has contributed to deeper inquiry as to why evolution might "choose" a particular suite of 20 amino acids from a much larger pool of plausible alternatives. Once again, the startling finding is just how clearly biology choice distinguishes from any reasonable expectation of random chance (Freeland & Philip, 2011).

From our perspective, efforts to understand the evolution of genetic coding by identifying adaptations exhibit clear conceptual parallels to the foundational work of linguists who have identified similar

explanations for the evolution of phonemic alphabets in vowel systems (Liljencrants & Lindblom, 1972; Lindblom, 1986). The primary difference seems to be one of empirical evidence: the recognition awarded to non-standard genetic codes is not yet matched by widespread acceptance of variations to "universal grammar." Cogent discussion elsewhere argues why this absence of evidence is entirely unsatisfactory both philosophically and scientifically as a justification for avoiding the question of where such a universal grammar has come from (MacNeilage, 2008: Chapter 1, pages 3-28), and we hope this is one area where the discoveries concerning the evolution of biochemical "language" will encourage those who persevere with evolutionary investigations to the acquisition of human language. Indeed, much of MacNeilage's (2008) approach to studying the mechanistic origins of human language parallels the other major approach to deriving clues to the history of genetic coding – the search for non-randomness that makes no sense in terms of adaptation. Here the undoubted pioneer of such thinking is Wong, who has worked for more than three decades to show that the standard genetic code can be traced back to earlier, simpler ancestral forms (which encoded fewer amino acids). A cornerstone of this argument is that only around half of the 20 amino acids used in modern genetic coding were available to early life (contemporary data for this point of view reviewed in Higgs & Pudritz, 2009 and Cleaves, 2010): the rest were invented by early evolution and then incorporated into the genetic code. As Wong points out, the sequence of biochemical steps or 'biosynthetic pathways' by which living organisms still synthesize each of the 20 amino acids seems to reflect this process: amino acids which appeared late within genetic code evolution are still made as chemical adjustments to amino acids that were likely available to the origin of life(Wong, 1975). Moreover, it appears that late-arriving amino acids may have entered the genetic code by capturing a sub-set of the codons previously assigned to their "early" counterparts, leaving behind patterns that make no sense in terms of error minimization. Once again, there is much more to this argument than we have space to discuss – and we again refer readers to one of several reviews that Wong offers to his own work (Wong, 2003). The important point is simply that in addition to its adaptive features and direct-templating origins, the code seems to carry a non-adaptive footprint of its evolutionary history. In this light, we concur strongly with the view (McNeilage, 2008: Chapter 2, especially pages 50-58) that it is strange, to say the very least, to claim that because language is a complex phenomenon, it self-evidently offers no quarter for meaningful evolutionary investigation (e.g. Chomsky, 1988 p. 170; Chomsky 2000, p. 4; p 49).

Origins of a generative grammar?

A current, widespread view from biology is that even the origin of genetic coding is logically distinct from the origin of genetic information. In particular, contemporary scientific thought favours the notion of an "RNA world," according to which the central dogma of molecular biology (i.e. the world of genetically encoded proteins) was preceded by one in which a single type of molecule, RNA, fulfilled all biological functions.

The RNA-world hypothesis derives support from two major sources. The first of these began with a Nobel prize-winning demonstration that RNA can spontaneously fold into shapes that catalyze specific biochemical reactions (1989, prize in chemistry). Subsequently, the same technology that has produced evidence for direct templating between RNA and amino acids (section "Genetic coding and arbitrary signifiers") allowed other researchers to uncover an unsuspected wealth of protein-like activities for other RNA molecules (Atkins, Gesteland and Cech, 2011). They too can fold, according to their sequence, into complex shapes that catalyze specific chemical reactions – indeed RNA molecules that do are known as ribozymes to mark their functional equivalence to protein enzymes.

A second, complementary source of support for the RNA-world hypothesis can be traced back to those who first deciphered genetic coding, including (who else?) Francis Crick. Their insightful observation was that the key molecular components of the genetic code were in fact not proteins at all. For example, in the words of Crick(1966, page 7), "*It almost appears as if tRNA were Nature's attempt to make an RNA molecule play the role of a protein*" and "*It is tempting to wonder if … the primitive machinery had*

no protein at all and consisted entirely of RNA" (Crick, 1968, page 371) These observations combined with identification of other, non-informational roles played by nucleotides in modern metabolism to create a concept "molecular fossils" (White 1976): molecules found in living systems unchanged from their ancient evolutionary debut. Technological breakthroughs of the gene-sequencing culture have allowed scientists to study specific molecules with enough detail to build an evolutionary paradigm expressed by another language metaphor: that modern metabolism is best understood as an evolutionary palimpsest - a document on which one layer of writing has been largely erased to make way for a second, later inscription (Benner, Ellington and Tauer, 1986). According to this view, a primordial metabolism constructed with interacting RNA molecules has been largely 'over-written' by a secondary layer of protein molecules since the advent of genetic coding.

We resist this current, widespread view and instead assert that neither line of evidence can support the RNA-world hypothesis over its exact conceptual opposite: a primordial, self-replicating metabolism of proteins which later "invented" RNA as genetic material. Our (distinctly minority) view is most easily explained in terms of the palimpsest metaphor. For a real palimpsest, the only way in which a chronology of two written layers is with reference to a third layer: the parchment on which the writing appears. Since biology contains no analogue for this writing-surface, nothing can tell us which "layer" came first. In other words, the existence of RNA molecules performing protein-like roles could equally well indicate that RNA evolved to usurp functions previously given over to proteins. RNA would have been capable of escaping from the role of genetic information storage precisely because of its incidental capacity to fold, form shapes and catalyze chemical reactions.

We introduce this unusual argument because this is where we see the most immediate potential for linguistic theory to inform ideas of biochemical evolution. Within linguistic theory, we are led to understand that anyone who does ask questions about the origins of generative grammar perceives that this generative grammar clearly emerged as an evolutionary refinement of some earlier, heterogeneous mixture of sounds and gestures that lacked a coherent, standardized framework. We further understand that this view is supported by a well-established body of appropriate logic (Pagin). We doubt that many evolutionary biologists would fail to see the logical difficulties (teleology) of supposing that a regularized information-storage system evolved in order to be ready for the subsequent evolution of content when presented with that clarity of thinking. However, most evolutionary biology was derived to describe the "modern" world in which the generative grammar of nucleic acid genes already exists. Here, nucleotide sequences are indeed the legitimate focal point for understanding most evolutionary outcomes, precisely because they are the physical basis of heredity (e.g. see Dawkins 1976). The careless extrapolation is to assume that nucleotide sequences have always played this role – that the origin of genetic information is one and the same thing as the origin of nucleotide sequences. From beyond the traditional disciplinary boundaries of evolutionary biology, this assumption seems questionable. For example, chemists widely agree that even individual nucleotides (the building blocks of RNA), let alone sequences of them joined together, are harder to form under pre-biological conditions than almost any other biologically relevant molecule; they also agree that amino acids are amongst the easiest. Astronomy agrees, having found no trace of nucleotides beyond Earth, but plenty of evidence to suggest that amino acids (and other simple organic compounds) quite literal rain down on the surface of every planet in the cosmos (Sephton). Meanwhile, biochemistry has long known that aside from their modern role as building blocks for genes, nucleotides play many other non-informational roles that happen to show every conceivable hallmark of "molecular fossil" status (such storing energy that is subsequently released to power protein-based metabolism). These are at best mere hints for an alternative view of the origins of genetic information – but they are consistent with something deep that linguistic theory seems to have already learned: that generative grammar is a refinement of an earlier, less modular system. If our previous discussion of genetic code evolution is intended to encourage those who study the evolution of human language, then this final comment reverses the relationship: it seems relevant and important to ask our evolutionary linguist colleagues to guide us through their understanding of the emergence of generative grammar to see what clarification it can bring to the study of life's origins.

Concluding remarks

In this paper we provide a brief overview of three themes relating to the broad topic evolution of biochemical "language": the origin of genetic coding, the emergence of the standard (not universal!) genetic code, and the origin of genetic information. The sequence in which we introduce these topics is deliberately chosen to present a progression from vague (and perhaps careless) use of linguistic analogy into increasingly direct suggestions of meaningful similarities. We are fully cognizant that metaphors are not to be mistaken for literal truths simply because the results would be interesting – but we consider ongoing exploration to be important. At the very least, if all linguistic constructions are metaphors for the concepts they represent, then all scientific explanations carry with them the strengths and weaknesses inherent to metaphor. They can illuminate and clarify our thinking or mislead us with the incidental baggage they bring along. Greater awareness of this point alone cannot harm scientists in their thinking, but we believe that far more is at stake.

In our introduction we identified ourselves as astrobiologists: scientists who actively seek to place their research and thinking within a broader framework of thought that defines life's place within the cosmos. The inherent interdisciplinarity of astrobiology has already yielded profound and unexpected insights. Placing the findings of astronomy next to those of chemistry and evolutionary biology reveals that half of the amino acids at work in your genetic code were likely brought to Earth by comets (reviewed in Cleaves, 2010; Higgs and Pudritz 2009). Placing the findings of geochemistry and planetary science next to those of biology reveals that life has changed our planet every bit as much as our planet has defined the environment for life (for example, the existence of an oxygen-rich atmosphere is a product of early life, not a pre-requisite for life (Kasting 2001); as much as half of Earth's minerals owe their very existence to biology (Hazen 2010). Overarching such individual insights, an overwhelming majority of scientific findings within the past fifty years suggest that life should be more common within our galaxy than was previously thought. For example, the existence of an extra-solar planet (i.e. a solar system other than our own) was confirmed as recently as 1995; the number of exoplanets known to science has been jumping ever since (for a recent estimate, see Cassan, A., Kubas, D., Beaulieu, J.-P., Dominik, M., Horne, K., Greenhill, J., Wambsganss, J., Menzies, J. *et al* 2012); life on our planet seems to have started much earlier than was previously thought (Mojzsis, 2002), and life appears to tolerate a much larger range of conditions than anyone guessed (Satyanarayana et al. 2005). All of this places an increasing significance on the importance of understanding whether life is destined by physics to converge, through evolution, towards predictable outcomes. At present, it remains ambiguous whether and where the superficial similarities of human language and biochemical "language" indicate something fundamental about the universe – about the emergence of generative grammars (combinatorial representation systems), about the way in which symbolic coding can emerge, and what signs of their evolution they leave behind. This would do much to inform our expectations about what characteristics we might legitimately expect from an independent origin of life. More parochially, it offers the potential to understand more about the relative roles of chance versus inevitability in steps that are fundamental to our own existence.

Acknowledgements

We are extremely grateful to Bjorn Lindblom and Francisco Lacerda for hosting the conference that led to this manuscript, to Peter MacNeilage and Peter Pagin for patience in explaining to us the fundamental ideas of evolutionary thinking within linguistic theory. We also thank Karen Meech for granting us time to pursue this particular interdisciplinary conversation. Sara Walker, Andrew Pohorille and James Stephenson have encouraged us in an exploration of alternatives to the RNA world hypothesis, and the NASA Astrobiology Institute has offered persistent encouragement for all interdisciplinary interactions that bring new insights to the study of life's origins. This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA09DA77A issued through the Office of Space Science.

References

Armstrong, K. (1993). A History of God. (*Ballantine Books*). Chapter 6 “The God of the Philosophers”, pp. 170-208.

Barrell, B. G., Bankier, A. T. & Drouin, J. (1979). A different genetic code in human mitochondria *Nature* 282: pp. 189 – 194.

Benner, S., Ellington, A.D. & Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. *Proc. Nati. Acad. Sci. USA* 86: pp. 7054-7058.

Caporaso, J.G., Yarus, M. & Knight, R. (2005). Error Minimization and Coding Triplet/Binding Site Associations Are Independent Features of the Canonical Genetic Code. *J Mol Evol* 61:597–607.

Cassan, A., Kubas, D., Beaulieu, J.-P., Dominik, M., Horne, K., Greenhill, J., Wambsganss, J., Menzies, J. et al (2012). One or more bound planets per Milky Way star from microlensing observations. *Nature 481*: pp. 167–169.

Cho, E.J., Lee, J.W. & Ellington, A.D. (2009). Applications of Aptamers as Sensor. *Annual Review of Analytical Chemistry* 2: 241–64.

Chomsky, N. (1988). *Language and the problems of knowledge.* (Oxford: Oxford University Press.).

Chomsky, N. (2000). *The architecture of language*. (Oxford: Oxford University Press.).

Cleaves, H.J., II. (2010). The origin of the biologically coded amino acids. *J Theor Biol* 263:490–498.

Crick, F. (1966). Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* 19: 548–55.

Crick, F. (1966). The Genetic Code –Yesterday, Today, and Tomorrow. *Cold Spring Harbor Symposia on Quantitative Biology* 31: 3-9.

Crick, F. H. C. (1967). An Error in Model Building. *Nature* 213: page 798.

Crick, F.H., Griffith, J.S., Orgel, LE. (1957). Codes without commas. *Proc Natl Acad Sci U S A* 43:416-21.

Crick, F.H.C. (1958). On Protein Synthesis. *Symp. Soc. Exp. Biol.* 12: 139-163.

Crick, F.H.C. (1968) .The Origin of the Genetic Code, *J. Mol. Biol.* 38: 367-379.

Des Marais, D., Nuth, J.A., Allamandola, L.J., Boss, A.P., Farmer, J.D., Hoehler, T.M., Jakosky, B.M., Meadows, V.S., Pohorille, A., Runnegar, B. & Spormann, A.M. (2008). The NASA Astrobiology Roadmap, *Astrobiology*, 8(4): pp. 715-730.

Filkins, P. (2012). How translation shapes meaning. *The Chronicle of Higher Education*, 58(28) pp. B12-B16.

Fox, G.E. (2010). Origin and evolution of the ribosome. *Cold Spring Harbor Perspectives in Biology* 2: pp. 1 – 18.

Freeland, S. & Philip G.K., (2010). Did evolution select a non-random “alphabet” of amino acids? *Astrobiology* 11: pp. 235–240.

Freeland, S. (2007). Could an intelligent alien predict earth’s biochemistry? In *Fitness of the Cosmos for Life* (eds. J. Barrow, S. Conway-Morris and S.J. Freeland), Cambridge University Press, Cambridge.

Freeland, S.J. & Hurst, L. (2004). Evolution Encoded, *Scientific American* 290:84-91.

Freeland, S.J., Wu, T. & Keulmann, N. (2003). The case for an Error Minimizing Standard Genetic Code., *Origins of Life and Evolution of the Biosphere* 33: 457-477.

Gamow, G. (1954). Possible relation between deoxyribonucleic acid and protein structures. *Nature*, 173: p. 318.

Gould, S.J.(1989). *Wonderful Life: the Burgess Shale and the Nature of History*. (W.H. Norton and Company, New York and London).

Hayes, B. (1998). The invention of the genetic code. *American Scientist*, 86: pp. 8-14.

Hazen, R.M. (2010). Evolution of minerals. *Scientific American* 303: pp. 58-65.

Higgs, P.G. & Pudritz, R.E. (2009). A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* 9:483–490.

Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. (1958). A Soluble Ribonucleic Acid Intermediate in Protein Synthesis *J. Biol. Chem.* 231: pp. 241–257.

Hughes, S. (2001). Making dollars out of DNA. The first major patent in biotechnology and the commercialization of molecular biology, 1974-1980. *Isis; an international review devoted to the history of science and its cultural influences* 92: 541–575.

James F. Kasting (2001) The Rise of Atmospheric Oxygen. *Science* 293: pp. 819-820.

Knight, R.D., Freeland, S. J. & Landweber, L. F. (2001). "Rewiring the keyboard: evolvability of the genetic code. *Nature Reviews Genetics*, 2:49-58.

Koonin, E.V. & Novozhilov, A.S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*. 61:99-111.

Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48: pp 839-862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In *Experimental phonology* (eds. J.J. Ohala amd J.J. Jaeger), pp. 13-14. Orlando, FL: Academic Press.

MacNeilage, P.F. (2008). *The origin of speech*. Oxford, Oxford University Press: pp 1-28.

Maxam, A.M. & Gilbert W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74(2).

Maynard Smith, J. & Szathmáry E. (1995). *The Major Transitions in Evolution.* (Oxford, England: Oxford University Press).

Mojzsis, S.J. (2002) Origin of Life: The First Fossils. In *Encyclopedia of Evolution*, edited by M. Pagel: Oxford University Press, Oxford, pp. 843-845.

Morris, S. C. (2003). *Life's Solution: Inevitable humans in a lonely universe* (Cambridge University Press, Cambridge UK).

Munger, S. (1999). Bible, Babel and Babble: The Foundations of Bible Translation (*International Bible Society, Colorado Springs*).

Oxford English Dictionary, online edn. 2E: http://oxforddictionaries.com/definition/code.

Rahman, M., "History of Genetic Engineering" by: http://www.ehow.com/video_5112340_history-genetic-engineering.html.

RNA Worlds: From Life's Origins to Diversity in Gene Regulation (eds. Atkins J. F., Gesteland F. And Cech T.R. Cold Spring Harbor Laboratory Press, America).

Rocha, R., Pereira, P.J., Santos, M.A. & Macedo-Ribeiro, S. (2011). Unveiling the structural basis for translational ambiguity tolerance in a human fungal pathogen. *Proc Natl Acad Sci U S A* 108:14091-6.

Rudolf, J. C. (2011). Cracking the Genetic Code of 'Brown Tides, *New York Times*, February 22.

Satyanarayana, T., Raghukumar, C. & Shivaji, S. (2005). Extremophilic microbes: Diversity and perspectives. *Current Science* 89: 78–90.

Saussure, F. (1983). Course in General Linguistics. (Eds. Bally C. and Sechehaye A.: Trans. Roy Harris. La Salle, Illinois: Open Court).

Sonneborn, T.M. (1965). Degeneracy of the genetic code: extent, nature and genetic implications. (Academic Press, New York).

The Nobel Prize in Chemistry 1972. Nobelprize.org. 13 May 2012 http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1972.

The Nobel Prize in Chemistry 1989. Nobelprize.org. 13 May 2012 http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1989/.

The Nobel Prize in Physiology or Medicine 1962. Nobelprize.org. 13 May 2012 http://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/.

The Nobel Prize in Physiology or Medicine 1968. Nobelprize.org. 13 May 2012 http://www.nobelprize.org/nobel_prizes/medicine/laureates/1968/.

The Nobel Prize in Physiology or Medicine 1969. Nobelprize.org. 13 May http://www.nobelprize.org/nobel_prizes/medicine/laureates/1969/.

Tuerk, C. & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–51.

Welton, M.G.E. & Pelc, S.R. (1966). XXX *Nature* 209: page 868.

White, H.B. (1976). Coenzymes as Fossils of an Earlier Metabolic State *J.Mol.Evol.*7,101-104.

Williams, D., Fournier, G.P., Lapierre, P., Swithers, K.S., Green, A.G., Andam, C.P. & Gogarten, J.P. (2011). A rooted net of life. *Biol Direct.*

Wong JT. (2005). Coevolution theory of the genetic code at age thirty. *Bioessays*. 27:pp 416-25.

Wong, J. T.-F. (1975). A Co-evolution Theory of the Genetic Code, *Proc. Natl. Acad. Sci. USA* 72: pp 1909−1912.

Yarus, M., Widmann, J.J. & Knight, R. (2009). RNA-amino acid binding: a stereochemical era for the genetic code. *J Mol Evol.* 69:406-29.

Yuan, J., O'Donoghue, P., Ambrogelly, A., Gundllapalli, S., Sherrer, R.L., Palioura, S., Simonović, M. & Söll D. (2010). Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett.* 584: pp. 342-9.

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. (*Evolving Genes and Proteins* Bryson V. & Vogel H. J. (eds), Academic Press, New York and London).