# NATURAL SELECTION AND THE GENETICALLY ENCODED AMINO ACID ALPHABET

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MICROBIOLOGY

MAY 2013

By Melissa A Ilardo

Thesis Committee:

Stephen Freeland, Chairperson
Sean Callahan
Paul Patek

## Table of Contents


**Ch.1 The Evolution of the Standard Genetic Code ..... 5**
1.1 Introduction ..... 5
1.2 Genetic coding is an evolved phenomenon ..... 7
*1.2a Codon Capture/Codon Disappearance ..... 8*
*1.2b Ambiguous Intermediate ..... 9*
*1.2c Unassigned Codon ..... 9*
1.3 Addition of new amino acids ..... 10
1.4 Emergence of the standard genetic code ..... 12
*1.4a The Stereochemical Hypothesis ..... 12*
*1.4b The Biosynthesis (Coevolutionary) Hypothesis ..... 15*
*1.4c The Adaptive Hypothesis ..... 18*
*1.4d Integrating ideas for genetic code evolution ..... 20*

**Ch. 2 Evolution of the Amino Acid Alphabet ..... 22**
2.1 Introduction ..... 22
2.2 Prebiotic synthesis of amino acids ..... 22
2.3 Broadening the case for early versus late amino acids ..... 25
2.4 The late amino acids ..... 27
2.5 Amino acids beyond the standard genetic code ..... 28
2.6 Summary ..... 31

**Ch. 3. Amino Acid Chemistry Space ..... 33**
3.1 Introduction ..... 33
3.2 The concept and applications of chemistry space ..... 33
3.3 Applying chemistry space to the amino acid alphabet ..... 35
3.4 Defining amino acid chemistry space ..... 39
*3.4a Size of amino acids ..... 40*
*3.4b Measuring amino acid hydrophobicity ..... 41*
*3.4c Measuring amino acid charge ..... 43*
3.5 Using chemistry space to investigate the genetic code ..... 44
3.6 Summary ..... 49

**Chapter 4 ..... 50**
4.1 Introduction ..... 50
4.2 Previous Analysis of Amino Acid Chemistry Space ..... 50
4.3 Enlarging our view of the amino acid universe. ..... 55
4.4 Re-testing the adaptive properties of standard amino acid alphabet. ..... 56
4.5 Summary ..... 59

**Chapter 5 ..... 61**
5.1 Introduction ..... 61
5.2 Suggested future directions ..... 61
*5.2a Further chemoinformatics analysis of amino acid alphabet properties ..... 62*
*5.2b Bioinformatics analysis of amino acid biosynthesis pathways ..... 63*
*5.2c Protein-building implications of alternative amino acid alphabets ..... 65*
5.3 Astrobiological frontiers for understanding the amino acid alphabet ..... 67

*5.3a Amino acids on Comets* ..... *68*
*5.3b Amino acids and the search for life on Mars* ..... *71*

**Appendices** ..... **74**

**References** ..... **75**

## List of Figures

Figure 1.1 The origin and evolution of the standard genetic code. 6
Figure 1.2 Known variants of the genetic code. 8
Figure 1.3. The 'twenty-first' and 'twenty-second' amino acids. 11
Figure 1.4. An illustration of Gamow's theorized 'diamond code'. 13
Figure 1.5 Yarus' DRT (Direct RNA Template) three-stage model for a stereochemical origin to genetic coding. 14
Figure 1.6 An example of Wong's proposed cession of codons from product to precursor, as mirrored in modern biological processes. 16
Figure 1.7 The Coevolutionary Hypothesis. 18
Figure 2.1 The complex reality of amino acids. 29
Figure 3.1 An application of chemistry space. 34
Figure 3.2 A broader chemical context of amino acid structures beyond those found in the standard genetic code. 37
Figure 3.3. An illustration of the strong correlation between Van der Waals volume (see main text) and the AAindex measure 'Residue volume,' (Bigelow 1967). 40
Figure 3.4 Low consensus in hydrophobicity measures. 42
Figure 3.5 A simple visualization of amino acid chemistry space. 45
Figure 3.6 Calculating the expansion of the "late" amino acids from the cluster of "earlies." 46
Figure 3.7 The chemistry space of the early and late amino acids. 48
Figure 4.1 Defining coverage: examples of range and evenness for hypothetical amino acid sets. 52
Figure 4.2 Summary of the results of Philip and Freeland 2011. 54
Figure 4.3 A two-dimensional view of amino acid chemistry space plotted for 4,147 L-chiral α-amino acid structures of plausible relevance to genetic encoding using MOLGEN predictions for LogP (hydrophobicity) and molecular volume (size). 57
Figure 4.4 The coverage of the standard amino acid alphabet compared to a comprehensive background of possible isomers. 59
Figure 5.1 The mechanism of Strecker Synthesis of Amino acids. 69

## List of Tables

Table 2.1. Growing consensus on early and late amino acids. 25
Table 3.1 Summary of the diverse arguments for individual members of the standard amino acid alphabet from (Weber 1981). 38
Table 3.2 Ten publications reporting different measures of volume for the genetically encoded amino acids taken from the AAindex. 41

# Ch.1 The Evolution of the Standard Genetic Code

## 1.1 Introduction

The origin of genetic coding is arguably the single most important event in the advent of life on Earth. The genetic code is the interface between two fundamental chemical languages: nucleotides and amino acids. The emergence of this interface represents a key step in the transition from a pool of inert organic molecules into replicating, evolving systems that are undeniably alive. Genetic coding also represents one of biology's biggest puzzles. Through physics, chemistry and astronomy we have come to understand much about the origin and diversity of prebiotic molecules on the early Earth. From bioinformatics and evolution we are increasingly confident in our understanding of the Last Universal Common Ancestor (LUCA) of all life on earth. However, the events surrounding the emergence of the genetic code, the link between inert prebiotic molecules and LUCA, are full of mystery.

From a pool of available molecules, life ended up using four nucleotides and twenty amino acids to encode and build its proteins. By the time of LUCA, the process of protein translation was largely fixed in the form of the standard genetic code (Figure 1.2) (Wachtershauser 1998). Since the time of this single-celled progenote some three and a half billion years ago (Woese 1977), life has used this simple decoding system to adapt to such a broad range of environments that twenty-first century scientists struggle to find a pressure high enough or a temperature low enough in which life does not thrive (Zeng 2009, Bakerman 2008). Clearly the standard genetic code provides an excellent system with which to diversify and adapt.

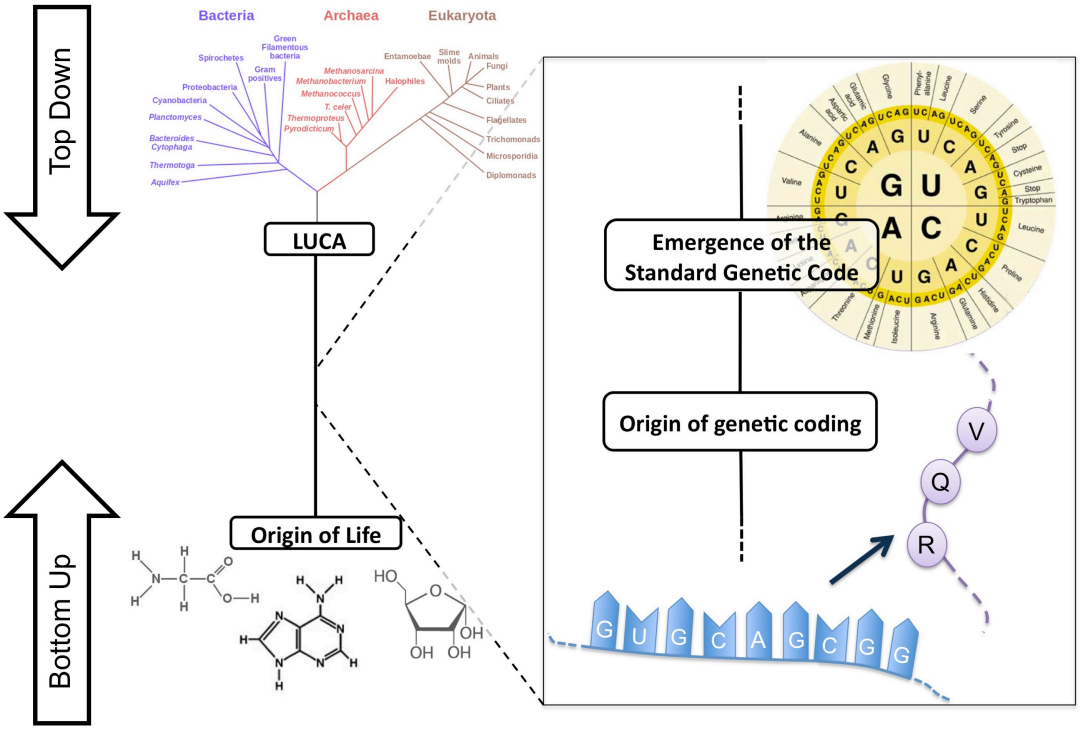



**Figure 1.1 The origin and evolution of the standard genetic code.**
The origin of genetic coding and the emergence of the standard genetic code represent a key connection between prebiotic molecules and LUCA (the Last Universal Common Ancestor), both of which are increasingly well understood. Studies of life's origins may be broadly divided into "bottom up" versus "top down" approaches. The former represent frontiers of astronomy, physics, geoscience and chemistry, which seek to understand the origin and distribution of key prebiotic molecules in and beyond the solar system. The latter represent research within the life sciences that seeks to work backwards from extant biology and biochemistry to understand life's earliest organisms. These approaches meet at the origin of life, which remains poorly understood and includes the origin of genetically encoded proteins and of the emergence of the standard genetic code.

New insights into the genetic code's seemingly remote evolutionary past not only help us to understand the deep history behind life on earth, but also seem likely to provide guidance as we move forward into a future of synthetic biology. Scientists in this emerging field are rapidly expanding the extent to which biological systems can be engineered from a "bottom-up" approach. In terms of the genetic code, this means that it is now relatively routine to insert unnatural amino acids into the genetic code of a model organism (reviewed in Liu 2010). Such innovation provides opportunities to explore new

proteins with novel functions and properties. Further understanding of the evolutionary factors that shaped natural genetic coding offer a well-defined context with which to gather insights and guidance for human-engineered genetic codes and the opportunities they imply.

As astrobiologists and evolutionary biologists delve deeper into life's origins and early evolution, and synthetic biologists develop technologies to bioengineer user-defined genetic codes, it is timely to review and extend current knowledge about how and why evolution produced the standard genetic code that forms the foundation of our present-day perception of biology.

## 1.2 Genetic coding is an evolved phenomenon

For decades after the discovery of the genetic code, the rules governing protein translation (i.e. which particular amino acid is encoded by each triple-nucleotide codon) appeared unchanged and unchanging across all life on Earth. Francis Crick presented as a possible reason for this universality the 'frozen accident' theory, which stated that any change in the way the code works would be lethal (or at least very strongly selected against) and therefore unlikely to become fixed in a population (Crick 1968). The simple logic was that, since the genetic code governs the translation of all genes into proteins, a change in the *encoding* of even a single amino acid would cause simultaneous errors in every protein containing that amino acid. Thus, while a fundamental tenet of evolutionary theory is that occasionally a mutation within a single gene may provide some adaptive advantage, any change to the rules by which all genes are translated would inevitably lead to disaster. However, in that same paper Crick alluded to suspicions about the possible existence of "ambiguous codons" that could represent more than one amino acid. The first confirmation for this characteristic prescience of Crick came more than a decade later when a non-canonical genetic code was found in our own species' mitochondrial DNA. Here, the codon UGA can be used to code for tryptophan rather than termination (Barrell et al 1979). The abundance of additional codon reassignments that have been found since then (see figure 1.2) suggests that the genetic code is not at all 'frozen', but rather has undergone and continues to undergo evolution (see Knight 2001 for a review).

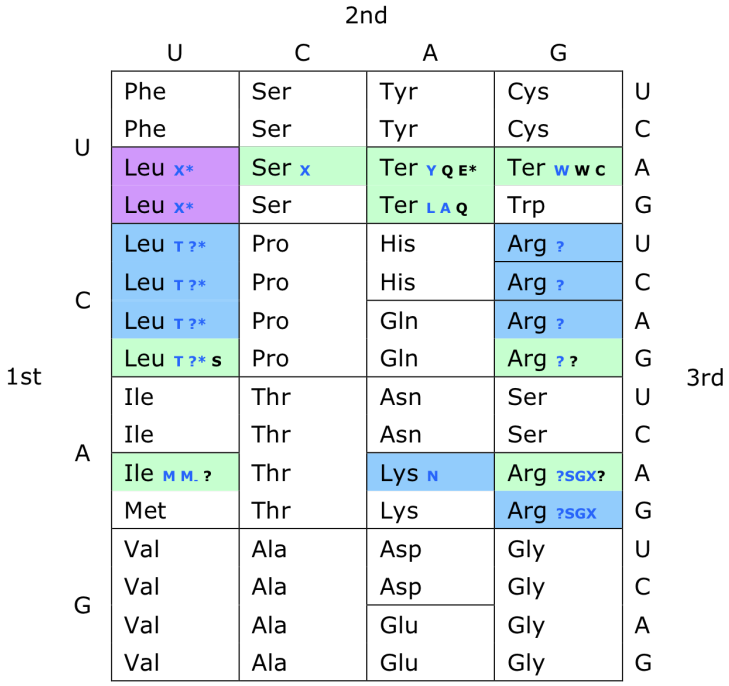



**Figure 1.2 Known variants of the genetic code.**
A subscript indicates particular codon reassignment, where blue letters represent changes in mitochondrial lineages, bold letters are changes in nuclear lineages, and starred letters indicate codon reassignments updated since Knight et al (2001). Blue boxes identify codons that have changed only in mitochondria, green boxes are codons that have changed both in mitochondrial and in nuclear lineages, and purple boxes show codons that have changed only in nuclear lineages. A negative subscript indicates reverse changes, and a ? indicates codons that have evolved to become unassigned (from Ilardo 2010).

There are currently three primary theories about the mechanisms of codon reassignment that help to explain how a series of mutations with deleterious intermediate stages can in fact become fixed within a population, circumventing Crick's persuasive logic.

### *1.2a Codon Capture/Codon Disappearance*

Codon capture is the only mechanism of reassignment in which a codon disappears entirely during the process. During evolution, events such as fluctuations in AG/GC pressure can cause a codon to be functionally replaced by one of its synonymous codons. For example, under a GC mutation bias codon UGU (Cys) might become replaced by

UGC (Cys). Further fluctuations may cause the codon to reappear in the genome, prompting rapid selection for the emergence of a new cognate tRNA. The corresponding new anticodon may be part of a tRNA molecule that is assigned to a different amino acid than the original tRNA (Osawa and Jukes 1989).

### *1.2b Ambiguous Intermediate*

In contrast to the central principle of codon capture theory, the ambiguous intermediate theory proposes an intermediate stage in which one codon is simultaneously recognized (decoded) by both a cognate tRNA and a near-cognate tRNA. The codon is driven out of this ambiguous state by mutations that either cause the cognate tRNA to become nonfunctional, improve the near-cognate tRNA's ability to read the codon, or some combination of the two. Once the near-cognate tRNA has a competitive advantage, it can replace the cognate tRNA as an act of natural selection, reassigning the codon to a new amino acid (Schultz and Yarus 1994).

### *1.2c Unassigned Codon*

In a nuanced blend of these two ideas, a third proposed mechanism is that mutations cause the primary tRNA associated with a codon to disappear, leaving the codon 'unassigned', though still translated (albeit inefficiently) by an alternative tRNA. This then prompts rapid evolution of a tRNA with greater degree of affinity for the codon to re-establish efficient translation (Sengupta and Higgs 2005, 2007).

These mechanisms can be unified by the gain-loss framework proposed by Sengupta and Higgs (2005). They describe a 'loss' as the deletion or loss of function of the gene that codes for the tRNA or release factors originally associated with a codon. A 'gain' occurs when a new type of tRNA becomes associated with the codon (i.e. it is reassigned) or when an existing synonymous tRNA mutates to pair more efficiently with the reassigned codon (Sengupta 2005). In whichever order these two events occur, the initial gain or loss may cause a selective disadvantage, however a subsequent,

compensatory loss or gain respectively creates a new code that eventually becomes selectively advantaged or neutral. In the case of codon capture, the complete disappearance of the codon has no selective disadvantage (Sengupta 2005).

For the purposes of this thesis, the detailed mechanisms of codon reassignment are less important than the underlying message: the pattern by which codons are assigned to amino acids can and does change over evolutionary timescales, even for the large and complex genomes of the current biosphere. This suggests caution in accepting any argument that the codon assignments of the standard genetic code represent a frozen accident of evolutionary history.

## 1.3 Addition of new amino acids

Selenocysteine (Sec) and Pyrrolysine (Pyl), the 'twenty-first' and 'twenty-second' proteinaceous amino acids, further enhance the view of the genetic code's evolutionary plasticity (Chambers 1986, Srinivasan 2002). Though both are built from modifications of conventional coded amino acids (Serine and Lysine, respectively), they are both legitimate additions to the amino acid alphabet.

The first of these to be discovered, selenocysteine (Sec), is used by representatives from all three domains of life (bacteria, archea and eukarya) (Hatfield 2002). As a molecule, selenocysteine may be thought of as a cysteine in which the sulfur group has been replaced by the more reactive counterpart, selenium. In the lineages where it occurs, this amino acid has captured the "stop" codon, UGA. Typical textbook accounts distinguish selenocysteine from the standard twenty amino acids by noting its translation is abnormal in three main aspects. First, the amino acid is not directly charged onto its appropriate $tRNA^{Sec}$, but rather is made *in situ* by an enzyme that produces selencoysteine from serine bound to the $tRNA^{Sec}$; second, the tRNA in question is unusual in structure and requires the presence of another protein (the SelB or mSelB) in order to compete its way into the ribosome for translation; third, in order for UGA to be translated as Sec, the mRNA in which its codon appears must also possess a second motif (the so-called SECIS element) that induces "misreading" of UGA from its usual meaning of "Stop" (Leinfelder 1988, Xu 2006, Yuan 2006, Commans 2006).

In fact, none of these objections makes Sec translation qualitatively different from "normal" translation of the standard 20 amino acids. Translation of glutamine and asparagine in some Archaeal lineages requires *in situ* enzyme modification of another amino acid bound to the appropriate tRNA (Feng 2004), all aminoacyl tRNA's use "elongation factor" proteins to complete translation (Daviter 2006), and recent literature has seen a growing number of reports that the effective translation of "normal" amino acids can depend on a broader mRNA context than the existence of three nucleotides listed as an appropriate codon in text-book illustrations of the standard genetic code (Moura 2011).

More recently, researchers have found a twenty-second amino acid, pyrrolysine (Pyl), encoded by the genomes of certain lineages of methanogenic bacteria within the domain archaea (Srinivasan 2002). Once again this added amino acid has taken over a stop codon, this time UAG. Furthermore, it is associated with a *cis*-acting motif (PYLIS) within the mRNA for effective translation (Krzycki 2005). In fact, its biggest difference from Sec-translation is that Pyl is charged directly onto its corresponding tRNA by a fairly normal-looking class II aminoacyl synthetase enzyme, which makes Pyl even less clearly distinguished than selenocysteine from the standard twenty amino acids.

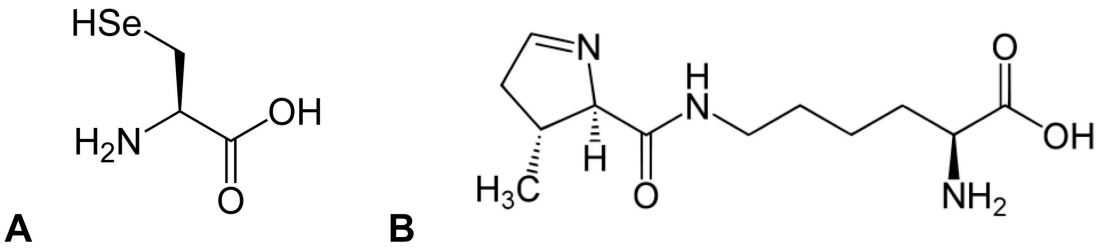



**Figure 1.3. The 'twenty-first' and 'twenty-second' amino acids.**
Chemical structures of selenocysteine (A) and pyrrolysine (B).

The genetic code is thus capable of expanding at the level of adding additional amino acids. As we shall see, this fits well with ideas about how the standard genetic code of 20 amino acids came into existence.

## 1.4 Emergence of the standard genetic code

Current science has advanced far beyond Crick's 'frozen accident' interpretation of the origin of the standard genetic code. Codon assignments can and do change, and new amino acids can be added to the code. Combined with the simple observation that the complex molecular machinery responsible for the standard code is a product of considerable evolution, it becomes legitimate and important to ask what else explains how and why one particular genetic code emerged within LUCA that still dominates the staggering diversity of life on our planet. Put another way, once we recognize the code as an evolvable phenomenon, we can ask what evolutionary forces shaped the emergence of the particular codon assignments found within the standard genetic code. Biological thinking has coalesced around three major ideas: the Adaptive Hypothesis, the Stereochemical Hypothesis, and the Biosynthetic or Co-Evolutionary Hypothesis.

### *1.4a The Stereochemical Hypothesis*

In some ways the simplest theory, the Stereochemical Hypothesis proposes that the genetic code results from direct chemical interactions between nucleotide sequences and amino acids. The famous physicist George Gamow initiated such thinking when he proposed a 'diamond code' which featured direct steric fit between genetic material and the amino acids into which it was translated (see figure 1.4). Although this neatly accounted for the code structure and codon assignments, it failed to take into consideration the (then unknown) adaptor molecule tRNA. Carl Woese further probed the stereochemical hypothesis in 1967 (just a year after launching his own alternative adaptive hypothesis, described below) by noting "*I am particularly struck by the difficulty of getting [the genetic code] started unless there is some basis in the specificity of interaction between nucleic acids and amino acids or polypeptide to build upon*" (Woese 1967).

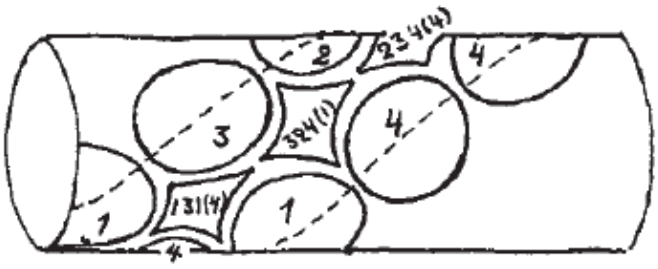



**Figure 1.4. An illustration of Gamow's theorized 'diamond code'.**
The amino acids (represented here as diamonds) neatly fit in the spaces between the three corresponding nucleotides (circles) by which they are encoded on the surface of a double-helix of DNA.

Early approaches to testing the logic of the stereochemical hypothesis were severely limited by the infancy of molecular biology. For example, in the same year as the complete structure of the genetic code was officially announced ("The Genetic Code" 1966 Cold Spr. Harb. Symp. Quant. Biol. 31: an entire, multi-authored volume dedicated to this landmark) Pelc and Welton utilized ball and stick models to show a direct steric fit between an amino acid and a codon (Pelc 1966). Unfortunately, within a year, Francis Crick identified an embarrassing error that entirely discredited their work: they had built their model backwards! While claiming to fit the amino acid Lysine with one of its codons AAG, they had instead shown that Lysine fit 'perfectly' with the codon GAA, which instead codes for Glutamic Acid (Crick 1967).

The lack of convincing empirical evidence for direct stereochemical interactions between codons and amino acids left a hole in such thinking that has only recently begun to be addressed through the advent of *in vitro* RNA biotechnologies. In particular, Michael Yarus and others have used SELEX (*in vitro* selection of initially random RNA fragment libraries) to find RNA molecules known as aptamers that target and bind a specific amino acid (Yarus 1988, Yarus 1991, Ellington 1990, Irvine 1990). Since the aptamers were selected from random libraries, any conserved sequences found in the end-products of selection imply functional importance – thus by sequencing, these researchers were able to identify sequence motifs of importance to amino acid binding. In this way, Yarus and colleagues have found that the codons and/or anticodons of arginine, glutamine, histidine, isoleucine, phenylalanine, tryptophan, and tyrosine were

significantly more likely to be found in conserved regions than non-conserved regions (reviewed in Yarus 2009).

It remains unclear why some of the aptamers that bind amino acids contain both codons *and* anticodons at higher frequencies than would be expected by chance. At first glance this seems reminiscent of the earlier conclusions of Pelc and Welton, which demonstrated that patterns and fits can be found anywhere if you look hard enough. Yarus explains this phenomenon, however, as the characteristic that allowed these particular amino acids to initiate the expansion from a Direct RNA Template (DRT), the earliest phase in his model of the genetic code's evolution, to the subsequent stage, a hemiDRT (see figure 1.5, Yarus 1998).

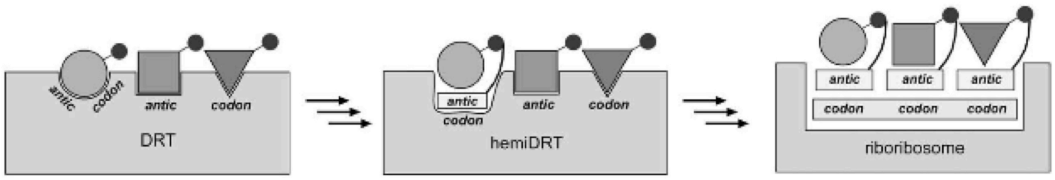



**Figure 1.5 Yarus' DRT (Direct RNA Template) three-stage model for a stereochemical origin to genetic coding.**
In the first panel, there is a direct steric fit between amino acids and their codons and/or anti-codons. The middle panel demonstrates a transitional partial direct template beginning to employ tRNA-like intermediates between the amino acids (shaded circle, square, and triangle) and the precursor to an mRNA sequence. In the final panel, we see the process as it occurs in modern organisms.

A relatively recent paper by Albert Erives has proposed a new stereochemical model for the origin of genetic coding that identifies stereochemical fits between an ancestral *proto-anti-codon* RNA (or pacRNA) and amino acids (Erives 2011). PacRNAs are theoretical hairpin tRNA precursors, an idea derived in earlier work (Hopfield 1978) and related to the more widely studied concept of proto tRNA half-mers (DiGiulio 2006, Fujishima 2008). This model neatly explains both the L-amino acid bias as well as the layout of the code, however it remains to be seen how robust the underlying chemical evidence is.

The stereochemical hypothesis's strength lies in its simplicity. The interactions it proposes are plausible, can explain the circumstances of code's origin, and support a

continuous transition from an ancient to a modern system. However, it is unclear how stereochemical interactions could be invoked to explain the expansion of the amino acid alphabet to include selenocysteine and pyrrolysine, as discussed in the previous section. These amino acids have clearly been added since the advent of both the ribosome and amino-acyl tRNA synthetase enzymes that charge tRNA species far away from any codon or anticodon. This raises an interesting question as to whether other amino acids within the standard genetic code were added after the advent of genetic coding, once stereochemical interactions were rendered unnecessary by the existence of tRNA adaptor molecules. This question takes on particular significance when we turn to consider the second major idea developed to explain the origin of the standard genetic code.

### *1.4b The Biosynthesis (Coevolutionary) Hypothesis*

The central concept of the second explanation for the emergence of the genetic code focuses upon amino acid pairs connected by metabolic pathways. In particular, proponents of this theory suggest that conserved pathways of amino acid biosynthesis observed in modern metabolism are historical remnants of the pathways and processes that first introduced new amino acids into the genetic code billions of years ago. Thus, it may be inferred that metabolic pathways reveal the history of genetic code evolution. The theory was first proposed in 1975 by J. Tze-Fei Wong under the name of genetic code "coevolution" (Wong 1975).

Wong described each of the twenty amino acids of the standard genetic code in terms of paired precursors and products. In each pair, the precursor amino acid is the substrate from which the product amino acid is made. He suggested that the set of precursors consisted of small, simple amino acids that are commonly believed to have been abundant in the early earth environment as a result of non-biological synthesis processes (an in-depth discussion of this statement will follow in Chapter 2). This sub-set was then expanded through the evolution of enzymatic pathways that created new 'product' amino acids.

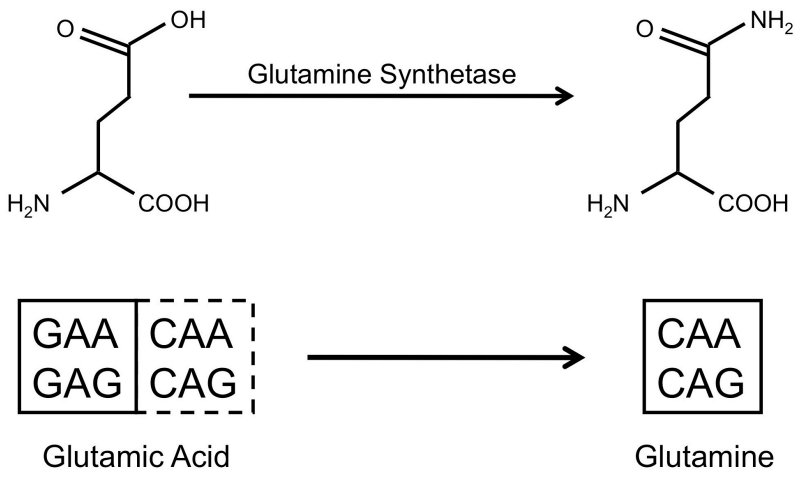



**Figure 1.6 An example of Wong's proposed cession of codons from product to precursor, as mirrored in modern biological processes.**
Glutamine synthetase produces glutamine as a metabolic derivative of glutamic acid. Within the standard genetic code, glutamic acid is encoded by codons GAA and GAG whereas glutamine is encoded by CAA and CAG. The "Coevolution" (or biosynthetic) hypothesis thus proposes that glutamic acid (a plausibly prebiotic amino acid) was originally encoded by 4 codons: GAA, GAG, CAA and CAG. When subsequent evolution "invented" glutamine as a biosynthetic derivative, it was incorporated into genetic coding by capturing two of the codons previously assigned to its precursor.

Although a roughly similar idea had been proposed by Dillon two years earlier (Dillon 1973), Wong's conceptual breakthrough was to argue that precursor amino acids would have ceded codons to their products. To take a specific example of the "coevolution" process, glutamic acid (Glu) is converted to glutamine (Gln) in modern organisms by glutamine synthetase (Figure 1.6). Under Wong's hypothesis, this biosynthetic pathway is a 'fossil' of a time when this process first introduced glutamine into the amino acid repertoire. At this time, precursor glutamic acid ceded two of its codons (CAA and CAG) to its product, glutamine, in order to produce the pattern of codon assignments we see in the standard genetic code (Figure 1.2). If the code emerged in this manner, then metabolic pathways of amino acid biosynthesis should have shaped the modern structure of the genetic code in a measurable way: precursor-product pairs should occupy contiguous (or touching) codon blocks with a high frequency. Wong observed that this is in fact the case. With few exceptions, product amino acid codon

domains are contiguous with those of their precursors (Figure 1.7a). Underlying aspects of this idea have gained considerable support from empirical observations, such as the fact that archaeal lineages modify Glu into Gln (and Asp into Asn) in just such a manner (Ibba 1997, Tumbula 2000). Indeed, a considerable literature has developed a variety of detailed models of this underlying idea; that amino acids sharing a biosynthetic pathway also tend to share nucleotide identity in the first base position of their codons (Taylor 1989, DiGiulio 2002).

The most notable criticism of the Biosynthetic theory as it was first proposed in 1975 is that many of the metabolic pathways on which the theory's claims depend seem to be much more variable in 2013 than they were when Wong first put forth his hypothesis (see Figure 1.7). In particular, a recent analysis of these pathways has revealed layers upon layers of additional complexity in the pathways by which various extant microbes interconvert the twenty amino acids of the standard genetic code (Hernandez-Montes 2007). Though many of Wong's initial pathways may exist within this updated view, it is a largely unexplored challenge to identify which pathways should be considered ancestral. One of the hallmarks of metabolism that Figure 1.7b makes clear is the constant overwriting of prior pathways. As far back as 1999, those interested in the early stages of biological evolution were noting that the correct interpretation of metabolic pathways might be "*hindered by … lateral transfers, replacements … and even of entire metabolic routes that may have been lost… it is possible there were alternative pathways which no longer exist or remain to be discovered*" (Lazcano 1999). Even if the pathways used by LUCA can be singled out, what, if anything, can we infer about the pathways of pre-LUCA that could have been representative of the code's origins, as Wong claims?

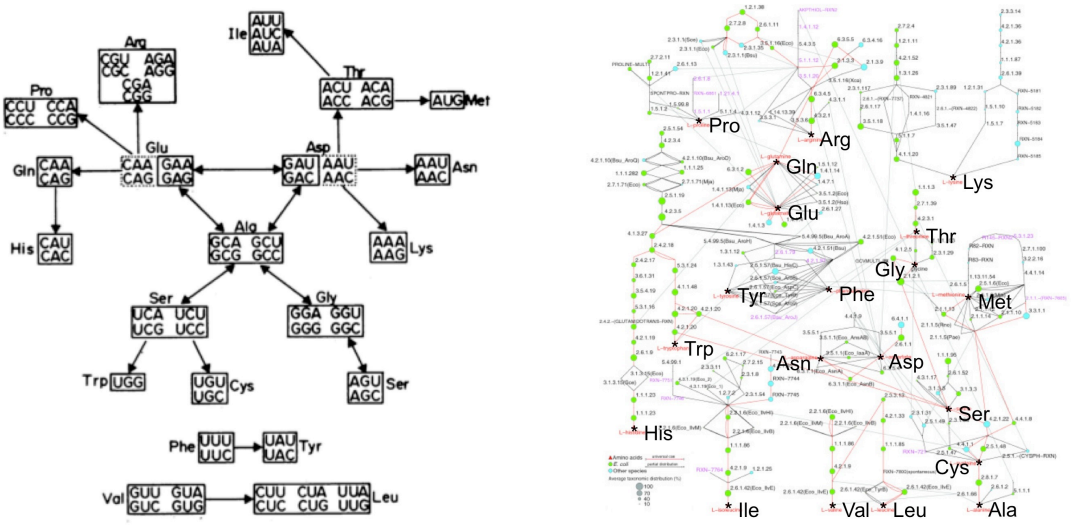



**Figure 1.7 The Coevolutionary Hypothesis.**
The pathways of amino acid biosynthesis from Wong 1975 [A] compared with the same pathways as we understand them today [B] (Wong 1975, Hernandez-Montes 2007).

## *1.4c The Adaptive Hypothesis.*

The third major idea advanced to explain the emergence of the standard genetic code, the Adaptive Hypothesis, postulates that the pattern with which 20 amino acids are assigned to 64 codons reflects natural selection to reduce the effect of genetic errors (point mutations, mis-transcription, and mis-translation). Before the complete structure of the genetic code was officially announced, researchers had already begun to notice that the distribution of codons to amino acids is decidedly non-random. In 1965, Zuckerkandl and Pauling (1965) as well as Sonneborn (1965) independently identified the error minimizing redundancy built into the code. Redundancy in this instance describes the property that a single amino acid is usually encoded by multiple codons. This reduces the potential for a point mutation to change the amino acid coded for by a given triplet by introducing the possibility of "silent" (synonymous) mutations. That same year, Carl Woese called attention to an additional, more subtle error minimizing property of the code: amino acids are often one point mutation away from those sharing similar physicochemical properties, such that the effects of genetic errors are lessened even when amino acid substitution takes place (Woese 1965).

The idea of adaptive redundancy lost momentum when in 1968 Francis Crick published the wobble hypothesis, which allows for non-standard base pairing between the first position of an anticodon and the corresponding third or 'wobble' position of a codon. The wobble hypothesis effectively dismissed the arrangement of the codons as a necessary result of tRNA's inability to discriminate between codons (Crick 1966). Such an intuitive explanation from a highly respected figure in biology halted debate within the biological community for a time about adaptive properties of the code. In particular, the subtler suggestion that similar amino acids were assigned to similar codons was largely overlooked. Although several rather abstract observations of this latter type were contributed from mathematicians and informational scientists (e.g. Cullman 1983, 1987; Figureau 1984, 1987, 1989), they did not penetrate mainstream thinking within biology.

In 1991 the adaptive theory was resurrected when Haig and Hurst (1991) published a quantitative test of Woese's underlying hypothesis that the standard codon assignments minimize the effects of mutations. Their work was an extension of an often-overlooked test performed over twenty years earlier (Alff-Steinberger 1969). As Alff-Steinberger had done before them, Haig and Hurst calculated the impact of mutations by quantifying the difference between amino acids before and after a single nucleotide mutation. Amino acids were determined to be more or less similar based on specific physicochemical properties. The 'error value', calculated as the average change in property for all possible point mutations, was measured for the standard genetic code and a large sample of computer-generated, randomized genetic codes. Both Haig and Hurst and Alff-Steinberger concluded that the code was unusual, but it wasn't until 1998 that the code was shown to be extraordinary when a repeat of the experiment, incorporated relative rates of transition and transversion mutations as well as mistranslational biases. This time, from a sample of one million hypothetical alternative codes, the standard genetic code proved to be the single most efficient at minimizing the effects of errors (Freeland 1998).

Critics and proponents of the adaptive code hypothesis alike have noted that these tests err towards oversimplicity. "*Evolutionary similarity of amino acids (meaning their substitutability within proteins) is unlikely to be perfectly represented by a single physiochemical measure… or by any simple combination of two or three such indices*."

(Freeland 2003). Substitutability of amino acids could alternatively be measured using PAM matrices that lie at the heart of sequence alignment algorithms (Sella 2006, Freeland 2000). However, this introduces circularity in the calculations. As one critic stated, "*since the PAM matrix counts the amino acid substitutions that occurred in families of homologous proteins during molecular evolution and as this process is mediated by the genetic code structure itself, it could be that the influence of the code on this matrix is such as to make any conclusion insignificant*" (DiGiulio 2001). In spite of this debate over potential biases, the extreme significance of the results suggests that a real effect is being measured.

### *1.4d Integrating ideas for genetic code evolution*

These three theories remain at the forefront of thinking about the origin of genetic coding, despite the fact that they are all decades old and have only received minor refinements since they were first proposed (e.g. compare Koonin 2009 with Knight 2001). This is not for lack of effort, including attempts that draw inspiration from as far left-field as super symmetry of quantum physics and p-Adic distances (Bashford 1997, Dragovich 2010). The strength and dominance over the years of the theories of stereochemistry, biosynthetic expansion, and adaptation suggest that further progress is more likely to be found by exploring the potential for integration of the three ideas rather than their replacement by fundamentally new concepts.

Qualitatively, it is easy to imagine how these hypotheses could overlap either in sequence or occur simultaneously (Knight 1999). For example, a strong stereochemical fit between an amino acid and an RNA adaptor molecule could have helped to establish that amino acid's use in a system of genetic coding. In the case of affinities between multiple amino acids and RNA motifs, natural selection could have played a role in determining which amino acid became incorporated on the basis of error minimization. Alternatively, a coding system launched through stereochemical affinities might also undergo re-shaping through adaptive codon reassignment and/or biosynthetic addition of new amino acids once tRNA adaptor molecules removed the need for direct steric interactions. This is especially interesting when we consider the secondary addition of

enzymatically-created novel amino acid variations to a simpler, earlier code: a process we see taking place with pyrrolysine and selenocysteine.

It is thus entirely possible (and in fact probable) that no one theory by itself wholly encompasses all of forces at work in shaping the origin and expansion of the standard code. Wong has acknowledged this fact and has gone so far as to estimate the relative weighting or influence of each of the three factors. He calculated that the relative contribution of each biosynthesis: adaptive: stereochemical influences to the selection of genetically encoded amino acids would be 40,000,000: 400: 1 respectively (Wong 2005).

The point of departure for this Masters' thesis, however, is to note that assessing the validity of all three theories (and any further estimation of their relative contributions) depends upon further investigations of two fundamental assumptions. These assumptions relate to the two previously mentioned chemical languages between which the genetic code acts as an interface: nucleotides and amino acids. A plethora of nucleotides and amino acids formed through biotic and abiotic processes were available in abundance during the earliest stages of life's evolution, as will be addressed in detail in Chapter 2. For the purpose of concluding this review of ideas regarding the evolution of the standard genetic code, what matters is to notice that any estimates made as to the relative importance of the theories described in this chapter build from the assumption of four nucleotides to encode twenty amino acids.

# Ch. 2 Evolution of the Amino Acid Alphabet

## 2.1 Introduction

The previous chapter built a case that the standard genetic code is a product of considerable biological evolution, and introduced the three major ideas that have sought to explain how and why one particular pattern of codon assignments emerged. It offered reasons for the plausibility of each hypothesis (the stereochemical hypothesis, biosynthetic hypothesis, and adaptive hypothesis) together with reasons why each should be regarded with some degree of caution. It concluded by noting that ongoing attempts to understand the relative contribution of each of idea to the emergence of the standard genetic code are flawed in so far as they treat the alphabet of 20 genetically encoded amino acids as a constant. This second chapter will clarify this final point by examining in greater detail the reasons to regard the amino acid alphabet as an evolved and evolvable phenomenon. The scope of this chapter is thus to review ideas and evidence for the evolution of the 20 amino acids of the standard genetic code.

## 2.2 Prebiotic synthesis of amino acids

In 1953, Stanley Miller performed a simple and elegant experiment that redefined concepts of prebiotic plausibility for fundamental biomolecules. Miller's experiment set out to test early theory regarding the formation of organic molecules in a primitive Earth ocean under an atmosphere of methane, ammonia, water, and hydrogen (Oparin 1938, Urey 1952, Bernal 1949). Miller's test circulated these compounds past an electric spark discharged within a sterile apparatus for several days. The resulting mixture was then examined using paper chromatography, which revealed the definitive presence of glycine, α-alanine, and β-alanine. There also appeared to be a weak signal from aspartic acid, though at the time this signature was considered inconclusive. Subsequent refinements of the experiment gradually increased the number of amino acids identified and by 1972

Miller's list of prebiotically plausible amino acids had grown to encompass nine (see table 2.1).

In 1979 Wong identified two limiting factors that could be restricting the number of amino acids identified through prebiotic simulations (Wong 1979). The first was the stability of the amino acids themselves. Wong hypothesized that the synthesis of asparagine and glutamine (chemical partners of aspartate and glutamate respectively) should be occurring under the conditions of the prebiotic simulation, and therefore their absence was due to the fact that they degraded before they could be identified. Wong's second theorized constraint was the state of the technology with which the amino acid composition of samples was being analyzed. Reanalysis of original Miller samples using modern instrumentation has since confirmed this latter factor (Parker 2011): paper chromatography indeed imposed a limit on the identification of amino acids. More importantly, these subsequent analyses have confirmed Wong's general underlying point: although a wide diversity of additional amino acids have been discovered, additional *genetically encoded* amino acids have not been found (Parker 2011). It therefore appears that despite improvements in instrumentation, a genuine asymptote occurs in terms of which genetically encoded amino acids can be produced through simulated recreation of early earth conditions.

A plateau in identification of proteinaceous amino acids within atmospheric simulation experiments supports the idea that a subset of the twenty genetically encoded amino acids, which we will refer to hereafter as the "early" amino acids, was prebiotically available. There is some room for debate about exactly where the cut-off for 'prebiotic plausiblility' occurs. For example, one thing that certainly could not be made using Miller's original technique was any amino acid containing a sulfur atom, as there was no sulfur present in the original mixture (Miller 1953). Thus methionine and cysteine (as well as non-proteinaceous, sulfur-containing amino acids) could never appear prebiotically plausible by virtue of the experimental conditions. Indeed, Miller performed a later experiment to remedy this limitation by adding $H_2S$ to the spark-tube mixture, though in his lifetime never analyzed the resulting test tube. A recent analysis of his original experimental output, a test-tube that had remained in cold storage for decades, showed that experiment had in fact produced six additional sulfur-containing

amino acids not reported by Miller including 5 non-proteinaceous amino acids (eg homocysteic acid) and, most important from the perspective of the standard genetic code, methionine (Parker 2011). Cysteine, the only other standard amino acid containing sulfur, was notably absent.

Since the 1950's, planetary science has reached a consensus that the atmospheric simulation experiments and their findings may not be representative of true prebiotic chemistry because Miller's 'strongly reducing' conditions are not characteristic of the early Earth (Kasting 1993, Cleaves 2008). Chapter 5 discusses this point in greater detail: for our current purposes this point is rendered moot for a variety of reasons. For example, the amino acids found by Miller are in remarkable agreement with the amino acids geologists have identified in carbonaceous chondritic meteorites such as the Murchison meteorite, which fell in Australia in 1969 (Kvenvolden 1970). Indeed, the convergence of results between amino acids identified in meteorites and prebiotic simulations was a major factor in Wong's proposal of an asymptote for prebiotic synthesis (Wong, 1979). The particular class of meteorites to which Murchison belongs contains such a wealth of organic compounds because they formed from the accretion of dust in the early solar system and were not subsequently subjected to high temperatures that would have disrupted their geochemical composition (Sephton, 2002). It is important to note that the Murchison meteorite and the amino acids it contains are considered representative of not only carbonaceous chondrites, but fundamental to understanding abiotic chemistry. As noted by Sandra Pizarello, a pioneer of amino acid analysis within meteorites, their study "*has long been part of investigations and discussions about the origin of life for the reason that [they] provide a natural sample of abiotic organic chemistry, and may offer insights on the possible environments and physico-chemical processes that fostered biogenesis … geological and biological processes of over four billion years have long eradicated any traces of early Earth's chemistry*" (Pizzarello 2010).

More broadly, it makes sense that a common inventory of amino acids are being found both in meteorites and in atmospheric simulations when the energy costs of amino acid production are taken into consideration. It seems clear that the early amino acids are the easiest and 'cheapest' to produce. In particular, in 1998 Amend and Shock calculated

the free energy of formation of amino acids from $CO_2$, $NH_4^+$, and $H_2$ in surface seawater conditions (Amend 1998). A recent meta-analysis compared these calculations with relative rankings based on availability of amino acids in a variety of prebiotic contexts including meteorites, hydrothermal synthesis, and atmospheric synthesis (Higgs 2009). This meta-analysis found a strong correlation between these measures ($r = 0.96$) for ten of the 'early' amino acids (in this case, G, A, D, E, V, S, I, L, P, T). Alternatively, if one simply considers the number of ATP molecules required to synthesize the amino acids using the biochemical pathways of *E.coli*, eight of the ten least costly amino acids are among the Miller 'earlies' (Akashi 2002). Interestingly, the two exceptions are N and Q, which Wong had previously highlighted as examples of amino acids that could be synthesized, but then rapidly degraded under the presumed prebiotic conditions. In other words energetics, prebiotic simulations and meteorite analysis all converge on a similar, coherent picture of which amino acids were likely available to an origin of life (Table 2.1).

| Study | G | A | D | V | L | I | E | S | T | P | M | Q | N | H | K | R | F | Y | W | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Miller 1972 | x | x | x | x | x | x | x | x | x | | | | | | | | | | | |
| Parker 2011 | x | x | x | x | x | x | x | x | x | | x | | | | | | | | | |
| Wong 1979 | x | x | x | x | x | x | x | x | x | | x | | | | | | | | | |
| Kvenvolden 1970 | x | x | x | x | | | | | | x | | | | | | | | | | |
| Higgs 2009 | x | x | x | x | x | x | x | x | x | x | | | | | | | | | | |
| Akashi 2002 | x | x | x | x | | | x | x | x | x | | x | x | | | | | | | |
| Trifonov 2000* | x | x | x | x | x | x | x | x | x | x | | | | | | | | | | |

**Table 2.1. Growing consensus on early and late amino acids.**
Regardless of what approach is taken, it appears that a similar conclusion is reached about which amino acids were present on the early earth. *Trifonov unfiltered: see main text, section 2.3, for detailed explanation

## 2.3 Broadening the case for early versus late amino acids

The consensus view described above highlights the agreement of a select few researchers, but what can be said about the community as a whole? Edward Trifonov attempted to develop a chronology for the amino acid alphabet that reflected the collective agreement of "*all available knowledge and thoughts about origin and evolution of the genetic code*" (Trifonov 2000). The forty criteria he collected ranged

from Miller's electric discharge experiments to evolutionary distances between isoacceptor tRNAs, and represent the findings of roughly 50 peer-reviewed scientific publications. From this extensive body of research, Trifonov ranked the amino acids according to each criterion in order to assign them a score for their likely position within a chronology of amino acid alphabet evolution.

Several considerations suggest reasons why it might be unwise to weigh all of the 40 criteria equally, such as the potential overlap in methodology (e.g. different measures of amino acid complexity) and the tendency for multiple researchers to build on a particular line of reasoning. Furthermore, Trifonov's analysis went on to do several rounds of 'filtering' of his data with somewhat subjective criteria. However, the point of central interest to this chapter is that Trifonov's results (especially the un-filtered chronology) matches remarkably well the division of early and late amino acids introduced above in terms of prebiotic synthesis (Table 2.1, final row). This match is remarkable precisely because Miller-type experiments and meteorite analyses represent only two of the forty criteria used in the analysis (Trifonov 2000). Indeed, although scientific truth is not well characterized as a "democratic vote of research publications," I advance here the view that the true power of Tirofonov's results may result from a curious statistical phenomenon commonly called the wisdom of the crowds.

Pioneered by Francis Galton (famed statistician who was also a cousin of Charles Darwin), the wisdom of the crowds, or *vox populi,* was built from a contest at a country fair attended by Galton. At the fair, a prize was offered to the person from the crowd who could most accurately guess the weight of an ox (Galton 1907). Upon gathering and analyzing the guesses of the contestants, Galton found, much to his surprise, that the average of the estimates was almost exactly correct; closer than nearly every individual attempt. A parallel can be drawn with the Trifonov experiment, in which the un-filtered results would most closely represent the *vox populi* of the genetic code research community. Any additional screening would be akin to Galton reducing the votes of all of the farmers in the contest to a single representative, simply because their estimates were likely produced using similar methodologies.

This interpretation of Trifonov's results is strengthened by comparison with two other meta-analyses that have appeared more recently within the scientific literature. One,

mentioned already, has reviewed the prebiotic plausibility from the perspective of physics, specifically energetic of formation and stability (Higgs 2009). The other, Cleaves (2010), has reviewed the same concept from the perspective of chemistry. This work suggested that the *metabolically* least costly amino acids are also among the earliest. The noteworthy observation is that these two reviews (each considering multiple hypotheses and empirical results from different disciplinary perspectives) have arrived at a remarkably consistent and uniform list of "early" versus "late" amino acids. Thus, while passionate debate continues about the location, conditions and mechanism of life's origins, it seems that this diversity of opinion may be largely irrelevant to our broad understanding of which amino acids were present at the earliest stages of genetic coding.

## 2.4 The late amino acids

So far, the emphasis of this chapter has been to build the case that a relatively clearly defined subset of the 20 genetically encoded amino acids was available to the origins of genetic coding. A complementary idea of equal importance is that a clearly defined subset of the 20 genetically encoded amino acids was *not* prebiotically plausible and therefore must be regarded as inventions of biological evolution.

Miller, the ultimate proponent of prebiotic synthesis of the genetically encoded amino acids, noted in 1980 that there was no clearly established prebiotic synthesis for lysine, arginine, histidine, or cysteine (Miller 1980). This would imply that they were necessarily the outcome of biotic processes, only possible after the advent of metabolism. Consistent with this observation, none of these amino acids have been identified in carbonaceous chondritic meteorites or subsequent prebiotic simulation experiments.

The concept of late amino acids was introduced already in Chapter 1 in conjunction with one of the major explanations for the emergence of the standard genetic code. The biosynthetic theory proposes specifically that a subset of late amino acids, the "late" amino acids, were biosynthetic inventions of the smaller subset of prebiotic 'precursor' amino acids. It is important to note, however, that the biosynthetic theory of code evolution goes beyond the concept of "early" versus "late" amino acids by proposing two additional, logically independent ideas: (i) that current pathways of amino acid biosynthesis are essentially the same as those by which "late" amino acids were first derived; and (ii) that "early" amino acids ceded codons to the "late" amino acids derived

from them. Both of these additional ideas are questionable – but the underlying, simpler idea of "early" versus "late" amino acids has lasted the test of time remarkably well.

Dissenting points of view do exist. For example, one recent phylogenetic reconstruction of putative ancient proteins concluded that "*An observable shift in amino acid usage at … conserved positions likely provides an untapped window into the history of protein sequence space, allowing events of genetic code expansion to be identified*" and goes on to identify Cys, Glu, Phe, Ile, Lys, Val, Trp, and Tyr as recent additions to the genetic code (Fournier 2007). While this list shows some overlap with table 2.1, it also shows some anomalies: most studies have concluded that Glu, Ile and Val are "early" amino acids. Another study that employed a similar bioinformatics approach, estimating the composition of proteins in an ancestral genome on the basis of conserved residues in descendant sequences, reached a different conclusion (Brooks 2002). Of the nine amino acids they found to be ancestral (Ala, Asn, Asp, Gly, His, Ile, Ser, Thr, and Val), eight have been previously identified as likely candidates for early amino acids. Given the diversity of approaches that converge to produce table 2.1, and the variability that surrounds the few dissenting views, we regard that the burden of proof currently resides with alternative views to explain how and why the broad consensus is wrong.

## 2.5 Amino acids beyond the standard genetic code

The third and final emphasis of this chapter is to demonstrate that many amino acids that are plausible candidates for genetic coding exist beyond those found in the genetic code. These amino acids were present when genetic coding originated and continue to be involved in biotic processes in all life on earth as non-coded biosynthetic intermediates. Understanding the existence of these amino acids is just as important to our understanding of the genetic code, for as Einstein once said, "*We not only want to know* how *nature is (and* how *her transactions are carried through), but we also want to reach, if possible, a goal which may seem utopian and presumptuous, namely, to know* why nature is such and not otherwise" (Einstein 1929).

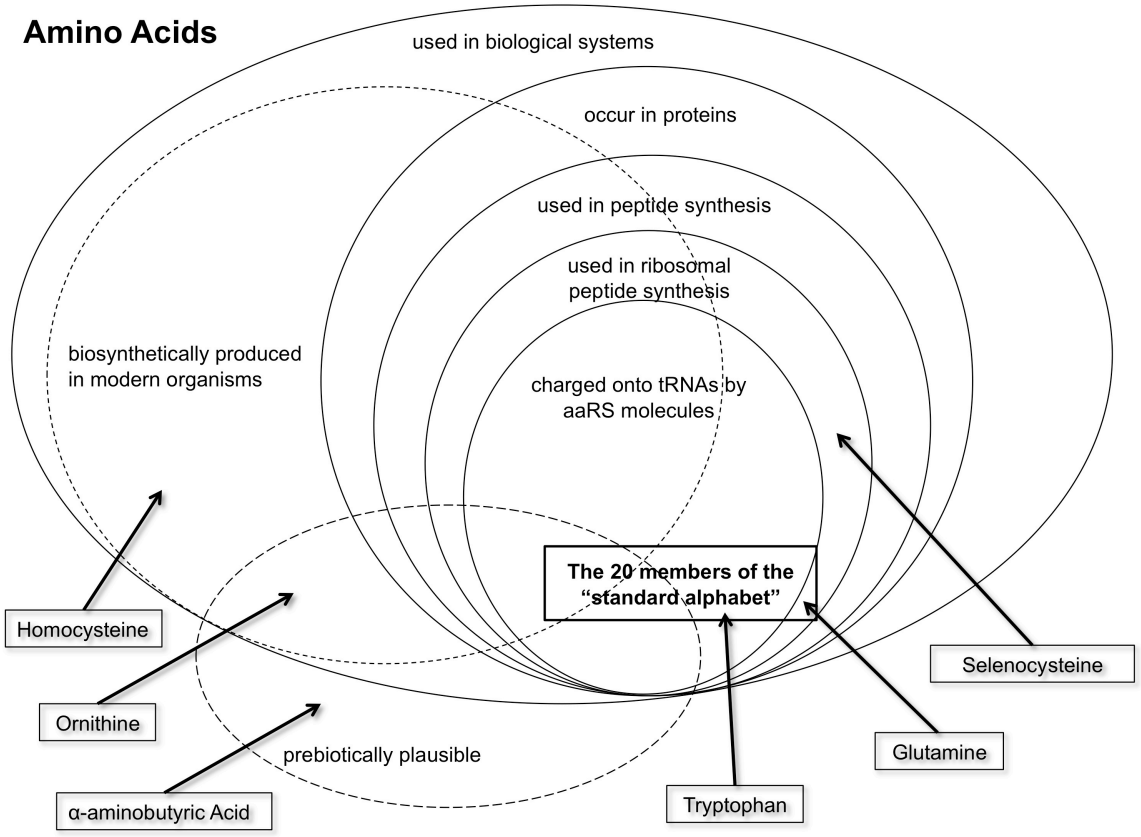



**Figure 2.1 The complex reality of amino acids.**
The twenty members of the standard amino acid alphabet represent only a small subset of those used in biological systems. Some examples of amino acids that lie around the edge of the standard twenty are also illustrated. Modified from (Freeland 2010).

Plausible abiotic, non-proteinaceous amino acids are abundant in prebiotic simulation experiments, with α-aminobutyric acid being tentatively identified as a component of Miller's earliest spark tube results. Twelve additional non-proteinaceous amino acids have since been identified in recent analysis of Millers original tubes (Parker 2011). Additionally, of the seventy amino acids found in the Murchison meteorite, only eight are members of the standard genetic genetic code (Cronin 1995).

In modern organisms, we find that even if we limit consideration to the metabolic pathways suggested by Wong as unchanged since the "invention" of late amino acids, then these pathways produce over a dozen additional, non-proteinaceous amino acids as biosynthetic intermediates. For example, homocysteine is an intermediate in the methionine cycle that can either be converted back to methionine or converted to cysteine via the transulfuration pathway. This amino acid and others like it must logically have

been presented as possibilities for genetically encoded amino acids as the code expanded, if the concept of biosynthetically derived 'lates' is true.

An interesting possibility allowed by this thinking is that the amino acid alphabet could even have been larger during some points in primordial evolution, and subsequently evolved simplicity by losing amino acids. This point was eventually conceded by Miller simply because of the abundance of non-coded amino acids relative to those that occur within the standard genetic code (Miller 1980), but the concept can be traced back earlier to Jukes (1973). He considered that the concept of evolutionary loss from the amino acid alphabet could explain the discrepancy between the number of codons assigned to Arg (6) and its surprisingly low usage in protein sequences. Specifically, he proposed that Arg was an 'intruder' that replaced its metabolic precursor, ornithine, by forming a stronger bond to ornithine's cognate tRNA.

Advances in synthetic biology show as an empirical fact that other amino acids (including those engineered by humans and those that could potentially exist as the result of (as yet) un-realized biosynthetic pathways) can be incorporated into genetic coding (Noren 1989). This plasticity of the current genetic coding machinery implies the possible existence of an earlier alphabet that used more amino acids than the current standard 20. An interesting suggestion for how the code might have arisen from a larger alphabet was presented by Fitch and Upper's "Ambiguity Reduction" hypothesis (Fitch 1987). They proposed three phases: an initial "fully ambiguous" genetic code, with no codon/amino acid pairings, a second phase of ambiguity reduction with coarse grain preferences (e.g. hydrophobic amino acids charged to acceptors recognizing a pyrimidine in the middle codon position and hydrophilic amino acids charged to acceptors recognizing a purine in the middle codon position), and finally the further reduction of ambiguity to the code as we know it today. Although the focus of "ambiguity reduction" was on the 20 amino acids of the standard genetic code, its essence is the notion that early, unsophisticated coding machinery decoded genetic messages according to broad physicochemical principles (e.g. translating a codon as "large, hydrophobic amino acid") rather than specific amino acid identities. This would fit well with the notion of a "fuzzy" primordial code that included amino acids beyond the standard twenty before evolving a specific set of couplings between codons and amino acids.

The major factor that currently limits investigations of a larger, earlier amino acid alphabet is the lack of any clear tests for such ideas: what would be the signature of amino acids that are no longer present within genetic coding, other than a general flexibility of the decoding machinery? For our purposes, it is sufficient to note that nothing we know argues against such an idea and, more positively, a wide variety of evidence suggests that many additional amino acids were available as possible alternatives to those that made their way into the standard genetic code.

## 2.6 Summary

This chapter builds a case for considering the 20 amino acids of the standard genetic code as comprising two different groups: "early" amino acids that were likely available at the origin for life through prebiotic syntheses, and "late" amino acids that are best understood as inventions of biology itself.

Under this view, it was only by constructing genetically encoded proteins made from early amino acids (perhaps augmented by ribozymes of a putative RNA-world) that life could have invented the late amino acids. If this view is correct, then it should be possible to construct enzymes made entirely from early amino acids that are capable of deriving the late amino acids. The plausibility of this idea is exemplified by pyrrolysine and selenocysteine, which are derived by enzymes constructed from the standard 20. That said, the exact chronology of amino acid alphabet evolution becomes important for any such models. For example, if tryptophan (Trp) were the 20$^{th}$ amino acid to be added to the genetic code, then Trp synthesis should be possible with an alphabet of the previous 19 amino acids. However, if Trp were incorporated as the 11$^{th}$ amino acid, then it should be possible to build enzyme(s) that can derive Trp using only the early 10.

While this hypothesis is testable in principle, it is not the subject of this thesis and is therefore outlined here as a possible area for fruitful further work in the future. For present purposes, the concept of a simpler, earlier system of genetic coding from which the standard genetic code evolved connects back to the theme of chapter 1: that the standard genetic code is an evolved and evolvable phenomenon. The amino acid alphabet at work in the standard genetic code, therefore, is best regarded as an evolutionary variable. The adaptive hypothesis therefore carries the implicit prediction that life's

"choice" of amino acids (as it "chose" from the pool of prebiotically plausible alternatives, and as it then expanded to the full repertoire through biosynthetic innovation) was influenced by selective pressures. That is, the set of genetically encoded early amino acids should show plausibly adaptive properties that distinguish them from random sampling of the products of prebiotic chemistry, and the late amino acids should have offered an adaptive advantage over the simpler, more limited amino acid alphabet.

# Ch. 3. Amino Acid Chemistry Space

## 3.1 Introduction

The first chapter of this thesis described how the codon assignments of the standard genetic code are thought to have emerged, creating an interface between genes and proteins through the interplay of three evolutionary factors: stereochemical interactions between nucleotides and amino acids, biosynthetic expansion of the amino acid alphabet, and natural selection for a code that minimizes genetic errors. It concluded by pointing out that further attempts to synthesize these lines of research will require greater recognition that the fundamental components of genetic coding are themselves evolutionary outcomes – i.e. variables that evolved to take specific parameters. Chapter 2 focused upon one of these variables, namely the amino acid alphabet, highlighting reasons to believe that the 20 amino acids of the standard genetic code divide into two groups, "earlies" and "lates" based on prebiotic plausibility. The chapter concluded by outlining how the adaptive hypothesis for the emergence of the standard genetic code would apply to this division: an adaptive explanation would posit that natural selection shaped the set of amino acids added to the standard genetic code.

The purpose of this third chapter is to introduce a methodological and conceptual framework that may be applied to the introductory material in order to develop testable hypotheses for the evolution of the amino acid alphabet.

## 3.2 The concept and applications of chemistry space

Researchers working at the interface of biology and chemistry have long recognized that "*the chemical compounds used by biological systems represent a staggeringly small fraction of the total possible number of small carbon-based compounds with molecular masses in the same range as those of living systems*" (Dobson 2004). This potential chemical diversity is foundational to drug discovery research, where scientists work to identify compounds that interact with human physiology in a beneficial manner, often by inhibiting or otherwise interfering with pathogenic

molecules. In recent years, a simple concept known as chemistry space has revolutionized the search for these bioactive molecules (e.g. Barker 2013, Lloyd 2006, Reymond 2012).

The fundamental point of chemistry space is to represent molecules with specific numeric values that define some aspect of their physical and/or chemical attributes (i.e. their physicochemistry). To do so requires the researcher to replace conceptual *properties* of interest (such as "hydrophobicity"), with precisely defined, measurable molecular *descriptors* (e.g. LogP – the partition coefficient of that molecule in a 2-phase system of water and octanol). This simple step transforms a collection of unique molecules into a precisely-defined set of points that exist within a single, multi-dimensional space amenable to quantitative analysis.

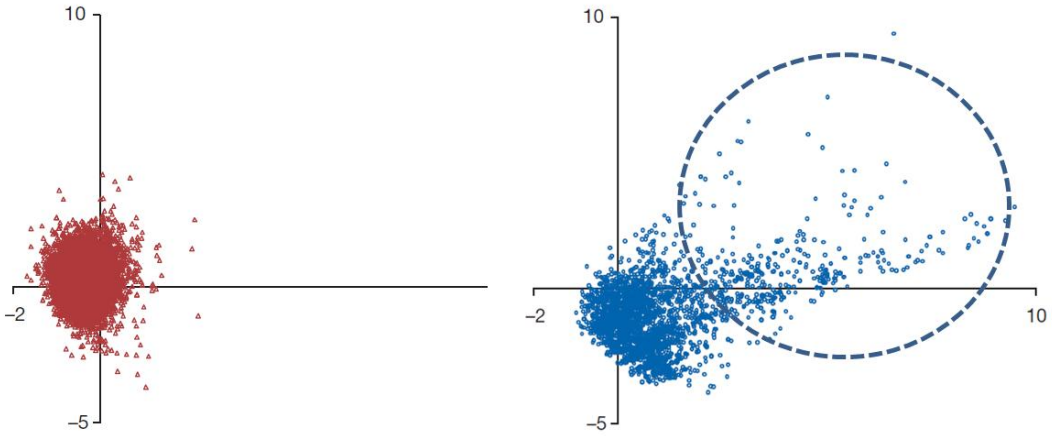



**Figure 3.1 An application of chemistry space.**
Comparison of the properties of (a) a large database of compounds created by combinatorial chemistry; with (b) natural products of known bioactivity. To visualize the similarities and differences of these molecules, principle components analysis was used to reduce a wide variety of quantifiable properties to 2 dimensions (which account for ~54% of total variation). The result is to reveal that natural products show far greater variation than the products of combinatorial chemistry. In particular, the area emphasized by a dashed ellipse indicates a region of bioactive natural products that is not represented by combinatorial chemistry, and thus represents fertile territory for future pharmaceutical research. This figure is adapted from Fig. 2 of Dobson (2004).

The significance of this transformation is to render conceptual questions about bioactive molecules in a framework suitable for advancing testable hypotheses. For example, if a collection of synthetic bioactive molecules and a comparable collection of natural bioactive molecules are each reduced to a 2-dimensional spread of their

quantifiable properties (Figure 3.1), then it becomes simple and intuitive to see that current efforts of combinatorial chemistry have overlooked significant sub-regions of chemistry space where undiscovered molecules of probable bioactivity lurk. In this case, the outcome of constructing an appropriate chemistry space is to suggest targets for further analysis and synthesis in the quest to find the next generation of pharmaceuticals. Implicit within this example is a subtle, important point: the definition of an appropriate chemistry space paves the way for scientific investigations of theoretical molecules that biology has yet to encounter. As long as the theoretical molecules may be described in terms of the same molecular descriptors as their natural counterparts, they become tractable components of the analysis. Deeper probing of such molecules is crucial both to enhance our understanding of the fundamental processes of life and to develop new strategies for treating disease.

A significant factor in the rising popularity of the chemistry space concept is that emerging tools of chemoinformatics are able to predict with reasonable accuracy the physciochemical properties of molecules that have been constructed by a computer (rather than synthesized in a laboratory). Appropriate computing can therefore create large combinatorial libraries comprising molecules of potential interest, and then predict their properties in order to render them as points within a suitable chemistry space (as shown in figure 3.1). This offers enormous savings in time and money from traditional, laboratory-based approaches.

## 3.3 Applying chemistry space to the amino acid alphabet

Amino acids provide an excellent example of the tiny subset of chemical possibilities explored by biology. The standard genetic code uses an alphabet of 20 L-chiral α-amino acids. Chemically, amino acids are simply monomers that can be linked together to form bioactive polymers. The enormous diversity of functional ribozymes (enzymes made from RNA) produced in little more than a decade of research offers good reason to think that fundamentally different building blocks could function in such a role. Returning closer to the familiar amino acids, chemical considerations suggest that the "backbone" consisting of a carbon atom linked at one end to an amine group and at the

other to a carboxylic acid, could easily have been a diamine or dicarboxylic acid instead (Weber 1981). Even if we limit our view to amino acids themselves, there is still enormous variety in the backbone that could function for building protein-equivalents. Genetic coding uses α-amino acids, where the α- prefix refers to the presence of a single carbon atom between the amine and carboxyl functional groups. However, β and γ-amino acids (amino acids with 2 and 3 carbon atoms, respectively) are also produced in the same prebiotic syntheses and meteorite analyses that are noted for containing "early" members of the standard alphabet. For these larger amino acid backbones, a side chain could attach at the β (or γ) carbon, rather than α (Figure 3.2). There is even the possibility that multiple side-chains could exist within a single monomer, limited only by the steric constraints of atomic crowding .

Surprisingly little scientific literature has dealt with this broader molecular context. Indeed the only extensive and focused discussion of the topic comes from a single publication by two synthetic chemists who spent their careers exploring the interface of chemistry and biology (Weber 1981). They noted, for example, that hydroxy acids are likely products of prebiotic synthesis, but polymerize to form polyesters, which (unlike polypeptides) are relatively unstable to basic hydrolysis. Furthermore, since the linkage between esters is not planar, polyesters cannot form intra-strand hydrogen bonds. Thus, "*some structures available to peptides would not form*" (Weber 1981). Similar chemical intuition led them to argue that amino acids were likely used instead of diamines or dicarboxylic acids due to the simple head-tail mechanism with which amino acids polymerize, whereby "*each addition of a new monomeric unit to the growing end yields a new terminus that possesses the same functional group as the previous terminus."* This consistent chemical structure at the growing end of a nascent protein is "*ideally suited to enzymatic catalysis*" as every step in polymerization requires only one, specific action – the formation of a peptide bond.

In regards to why genes encode α (and only α) amino acids, Weber and Miller noted that an early translation system using a mixture of backbones (i.e. polymerizing α, β , γ, and δ amino acids), would have likely required a greater diversity of enzymes to accurately polymerize each possible combination (i.e. one enzyme to polymerize α with α, another for α with β , etc.). More generally and perhaps more compellingly, they noted

that secondary structure with non-α amino acids is rare or entirely absent, which may be attributed to "*the polymer's entropy that is caused by internal rotation about the $C^{\alpha}$-$C^{\beta}$ bond*…." (Glickson 1971, Balasubramanian 1974) (see figure 3.2).

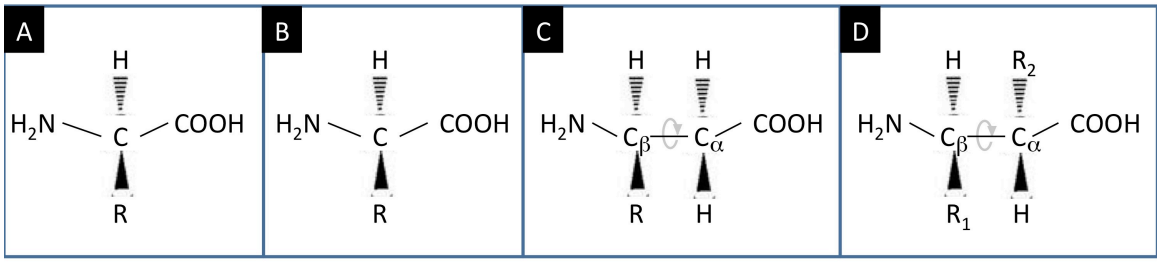



**Figure 3.2 A broader chemical context of amino acid structures beyond those found in the standard genetic code.**
(A) The 20 amino acids of the standard genetic code are all L-chiral α-amino acids in which all molecular variation is limited to the side-chain (R); (B) all indications are that R-chiral α-amino acids, stereoisomers of those found in the genetic code, would function equally well in a "mirrored" world of molecular biology; (C) β-amino acids contain an additional (β) carbon in their backbone, leading to an additional degree of rotation which prevents any easy formation of secondary or tertiary structure; (D) The addition of one or more additional carbons to the backbone of an amino acid opens up new possibilities for additional side-chains (e.g. R2).

For the purposes of this thesis, and to achieve a pragmatically tight focus for research, we choose to accept these arguments at face value while noting only the relative dearth of scientific literature that has corroborated or challenged the arguments put forth by Weber and Miller. By analogy to Crick's frozen accident explanation for the standard genetic code, we suggest that it may be premature to consider that science understands why exactly L- α-amino acids are nature's "choice" on the basis of compelling, logical arguments that have not been tested. With that caveat, we focus instead only on the variety of side-chains that can exist on an L-chiral α-amino acid backbone. Even with this narrow focus, the first point to make is the relative lack of systematic scientific research to explore the structural diversity that is possible, let alone the corresponding properties of these molecules. The side chains of the standard amino acid alphabet represent only a small sample of many permutations for the same underlying set of atoms. Thus, even limiting consideration to amino acids with the same backbone as those found in biology, as many or fewer carbon atoms in their side chains, and the same functional groups present in their side-chains, we still face a vast and largely uncharted chemistry space!

| Amino Acid | Notes from Weber and Miller (1981) | |
|---|---|---|
| *Hydrophobic Amino Acids* | | |
| Glycine | | Abundant in both electric discharge experiments and the Murchison meteorite |
| Alanine | | |
| Valine | | |
| Leucine | | |
| Isoleucine | | |
| α-amino-*n*-butyric acid | Surprising in their absence from the standard alphabet | |
| norvaline | | |
| norleucine | Possibly absent due to structural similarity with methionine | |
| *Cyclic amino acids* | | |
| Proline | Proline's advantage over pipecolic acid is the rigidity of tis ring, which would be less structurally flexible in proteins | |
| Pipecolic Acid | | |
| *Acidic amino acids* | | |
| Aspartic acid | Formed in prebiotic syntheses and found in Murchison, these are "*very logical… assuming acidic amino acids are needed*." | |
| Glutamic acid | | |
| *Basic amino acids* | | |
| Lysine | No clear prebiotic synthesis, "*free amino groups are needed*" | |
| Arginine | No clear prebiotic synthesis, likely present due to its permanently charged side chain, shortest guanidine amino acid suitable for proteins | |
| Histidine | No established prebiotic synthesis, "*most suitable imidazole containing amino acid*" | |
| *Hydroxy amino acids* | | |
| Threonine | | Simplest hydroxy amino acids |
| Serine | Found in electric discharge products but not Murchison, degrades quickly | |
| Homoserine | Disadvantage is rapid lactonization of activated esters | |
| Cysteine | Most likely thiol amino acid, hasn't been found in Murchison or electric discharge products | |
| Methionine | Synthesized by electric discharge experiments via acrolein, instability is due to oxidation by $O_2$ | |
| *Aromatics* | | |
| Phenylalanine | Only prebiotically synthesized aromatic amino acids | |
| Tyrosine | | |
| Tryptophan | | |
| Asparagine | Likely late additions, stable in peptides | |
| Glutamine | | |

**Table 3.1 Summary of the diverse arguments for individual members of the standard amino acid alphabet from (Weber 1981).**

In the same paper that tackled the broader chemical possibilities and alternatives to amino acids, Weber and Miller went so far as to provide individual reasons why each amino acid was a likely candidate for incorporation into the genetic code, including consideration of those, such as alpha-amino-n-butyric acid, which are surprising in their absence from the standard amino acid alphabet. Their arguments built primarily from the concept of prebiotic plausibility, with secondary considerations of stability added in. A brief summary of these arguments is provided in table 3.1.

The central point of this chapter is to investigate whether an appropriately defined chemistry space can simplify this analysis. The principal challenge to testing this idea, therefore, is to identify the descriptors that most accurately reflect the biologically relevant chemistry space of the amino acids.

## 3.4 Defining amino acid chemistry space

Defining the chemistry space of amino acids is essential as it provides a quantitative method of analyzing the growth of the amino acid alphabet. However, it is challenging given the vast array of amino acid molecular properties that have been measured. The Amino Acid Index (or AAindex) comprises an extensive collection of such measures for the genetically encoded amino acids harvested from the scientific literature (Kawashima 1999). Currently, the database lists over 600 molecular descriptors. Though few of these descriptors are entirely independent of one another, the question remains: which subset best reflects relevance to the role of building proteins? To address this challenge, we note that three key properties are commonly acknowledged to determine amino acids' biochemical roles within protein structure and function: size, hydrophobicity, and charge (Grantham 1974, Ladugna 1997).

Each property contributes to the biochemical interactions of amino acids in unique and essential ways. Amino acids with similar sizes have been demonstrated to be highly exchangeable during protein evolution, indicating that size is an important contributor to defining amino acid similarity (Grantham 1974); hydrophobicity is widely acknowledged as a fundamental determinant of folding pathways of nascent peptides and has been previously linked to the genetic code through the adaptive hypothesis (as

discussed in Chapter 1) (Kauzmann 1955, Wrabl 2005, Freeland 1993); and electrostatic interactions between amino acids have been shown to play a crucial role in inter- and intra-molecular protein interactions (Gilson 1987). For these reasons, and because of the exhaustive tests that have been performed of the reliability of various measures of these properties (Lu 2006), these are the three dimensions we have chosen to investigate.

### *3.4a Size of amino acids*

Amino acid size is an intuitive concept, yet even for such a general property there exist a variety of potential descriptors, including measures of molecular weight, side chain length, and volume. For our purposes, we chose to measure amino acid size using volume. Previous research indicates that amino acid volume may be predicted with good accuracy using chemoinformatic software (Lu 2006). In addition, an examination of volume measures available through the AAindex indicates that precise choice of descriptor is unlikely to affect our view of chemistry space.

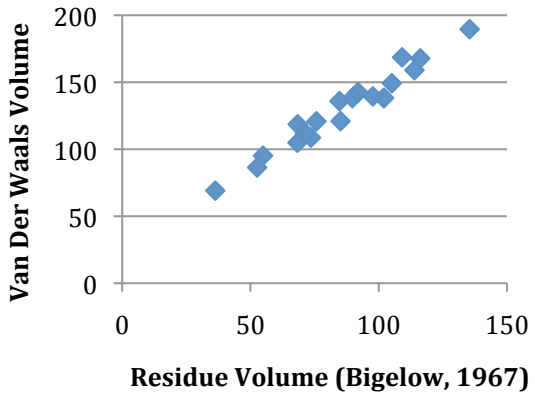



**Figure 3.3. An illustration of the strong correlation between Van der Waals volume (see main text) and the AAindex measure 'Residue volume,' (Bigelow 1967).**
Although this represents only one pair of volume measures, the volume values reported in all ten studies available in the AAindex (table 3.2) were significantly positively correlated with each other, and with our chosen measure of ACD molar volume, to a similar extent. This agreement illustrates a fundamental understanding of the conceptual property "size" and its measurement reflected in corresponding molecular descriptors.

Specifically, the Amino Acid Index lists 10 molecular descriptors that explicitly describe themselves as measures of volume (see table 3.2). A simple analysis finds that

all of them are significantly correlated with each other ($p < 0.01$[1]) and with an eleventh measure, Van der Waals volume, which we will use to describe the size of amino acids. This strong consensus between all available measures of volume indicates that different approaches have reached a fundamental understanding of how to meaningfully record property of amino acid size: all measures reflect the same underlying physicochemical property.

| Author | Year | Measure | AAindex ID |
|---|---|---|---|
| Bigelow | 1967 | Residue volume | BIGC670101 |
| Chothia | 1975 | Average volume of buried residue | CHOC750101 |
| Fauchere et al | 1988 | Normalized van der Waals volume | FAUJ880103 |
| Goldsack-Chalifoux | 1973 | Residue volume | GOLD730102 |
| Grantham | 1974 | Volume | GRAR740103 |
| Krigbaum-Komoriya | 1979 | Side chain volume | KRIW790103 |
| Tsai et al | 1999 | Volumes including the crystallographic waters using the ProtOr | TSAJ990101 |
| Tsai et al | 1999 | Volumes not including the crystallographic waters using the ProtOr | TSAJ990102 |
| Harpaz et al | 1994 | Mean volumes of resides buried in protein interiors | HARY940101 |
| Pontius et al | 1996 | Average volumes of residues | PONJ960101 |

**Table 3.2 Ten publications reporting different measures of volume for the genetically encoded amino acids taken from the AAindex.**
All of the available volume measures are significantly correlated with each other. This strong agreement suggests a meaningful consensus has been reached in measuring amino acid volume.

### *3.4b Measuring amino acid hydrophobicity*

In contrast to the strong consensus for different descriptors of amino acid size, hydrophobicity appears far less straightforward to measure. It is in fact an excellent example of the difference between a molecular descriptor and the abstract property it seeks to represent. Hydrophobicity is a real and tangible concept, widely recognized as a key to protein folding, yet recording hydrophobicity proves highly problematic to define as a descriptor.

[1] This probability was calculated using the Pearson product-moment correlation coefficient (Steel 1980).

Consensus on measuring hydrophobicity has not even extended to whether large numbers should indicate a preference for or against an aqueous environment. It is therefore perhaps not surprising that if we simply take the 26 AAindex studies that contain 'hydrophobicity' in their database entry title, then of the 325 possible pairings of any two studies, 59 were not significantly correlated with each other, and an additional 90 pairs were significantly *negatively* correlated (see figure 3.4). Beyond the simple conceptual question of which way a hydrophobicity scale should be ordered, a more subtle confounding factor is that meaningful correlation depends on linear scaling of the variables under consideration. For diverse attempts to measure hydrophobocity, we have no guarantee that this assumption holds: LogP is, as its name implies, a log scale. Other ways of measuring hydrophobicity offer no guarantee that the resulting scale will not form a curvilinear distribution (or any other conceivable mathematical function). Amongst other things, measuring a molecule's "fear of water" depends very much on what alternative solvent(s) are presented. Indeed, the abundance of insignificant correlations seen in figure 3.4 suggests that different descriptors of hydrophobicity correspond vaguely at best to a single underlying concept.

| | ARGP | CIDH | CIDH | CIDH | CIDH | CIDH | EISD | GOLD | JOND | PRAM | SWER | ZIMJ | PONP | WILM | WILM | WILM | WILM | JURD | WOLR | KIDA | COWR | BLAS | CASG | ENGD | FASG | GEOD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARGP820101 | | 0.75 | 0.86 | 0.83 | 0.82 | 0.87 | 0.55 | 0.94 | 1.00 | -0.42 | 0.73 | 0.75 | 0.60 | 0.66 | 0.63 | 0.18 | 0.36 | 0.49 | 0.35 | -0.71 | 0.68 | 0.76 | 0.59 | -0.42 | -0.60 | -0.57 |
| CIDH920101 | 0.75 | | 0.92 | 0.87 | 0.83 | 0.92 | 0.58 | 0.71 | 0.75 | -0.55 | 0.85 | 0.64 | 0.79 | 0.54 | 0.58 | 0.39 | 0.21 | 0.59 | 0.34 | -0.71 | 0.56 | 0.74 | 0.79 | -0.55 | -0.79 | -0.69 |
| CIDH920102 | 0.86 | 0.92 | | 0.91 | 0.90 | 0.97 | 0.61 | 0.83 | 0.86 | -0.57 | 0.87 | 0.62 | 0.82 | 0.64 | 0.64 | 0.28 | 0.29 | 0.61 | 0.37 | -0.76 | 0.66 | 0.81 | 0.80 | -0.57 | -0.79 | -0.74 |
| CIDH920103 | 0.83 | 0.87 | 0.91 | | 0.96 | 0.97 | 0.74 | 0.77 | 0.83 | -0.65 | 0.87 | 0.70 | 0.90 | 0.65 | 0.61 | 0.32 | 0.38 | 0.78 | 0.56 | -0.79 | 0.75 | 0.84 | 0.83 | -0.65 | -0.85 | -0.73 |
| CIDH920104 | 0.82 | 0.83 | 0.90 | 0.96 | | 0.97 | 0.78 | 0.78 | 0.82 | -0.71 | 0.86 | 0.65 | 0.93 | 0.73 | 0.71 | 0.33 | 0.49 | 0.83 | 0.61 | -0.84 | 0.80 | 0.88 | 0.90 | -0.71 | -0.90 | -0.80 |
| CIDH920105 | 0.87 | 0.92 | 0.97 | 0.97 | 0.97 | | 0.71 | 0.82 | 0.87 | -0.64 | 0.89 | 0.69 | 0.89 | 0.68 | 0.68 | 0.32 | 0.37 | 0.73 | 0.50 | -0.80 | 0.73 | 0.85 | 0.86 | -0.64 | -0.86 | -0.77 |
| EISD840101 | 0.55 | 0.58 | 0.61 | 0.74 | 0.78 | 0.71 | | 0.54 | 0.54 | -0.94 | 0.72 | 0.47 | 0.73 | 0.73 | 0.59 | 0.29 | 0.21 | 0.89 | 0.91 | -0.90 | 0.86 | 0.88 | 0.73 | -0.94 | -0.80 | -0.70 |
| GOLD730101 | 0.94 | 0.71 | 0.83 | 0.77 | 0.78 | 0.82 | 0.54 | | 0.94 | -0.46 | 0.79 | 0.77 | 0.59 | 0.75 | 0.66 | 0.05 | 0.36 | 0.50 | 0.38 | -0.66 | 0.69 | 0.82 | 0.52 | -0.46 | -0.55 | -0.62 |
| JOND750101 | 1.00 | 0.75 | 0.86 | 0.83 | 0.82 | 0.87 | 0.54 | 0.94 | | -0.42 | 0.73 | 0.75 | 0.60 | 0.66 | 0.63 | 0.18 | 0.36 | 0.48 | 0.35 | -0.71 | 0.68 | 0.76 | 0.59 | -0.42 | -0.60 | -0.57 |
| PRAM900101 | -0.42 | -0.55 | -0.57 | -0.65 | -0.71 | -0.64 | -0.94 | -0.46 | -0.42 | | -0.69 | -0.36 | -0.68 | -0.61 | -0.51 | -0.36 | -0.19 | -0.86 | -0.88 | 0.87 | -0.78 | -0.86 | -0.73 | 1.00 | 0.76 | 0.83 |
| SWER830101 | 0.73 | 0.85 | 0.87 | 0.87 | 0.86 | 0.89 | 0.72 | 0.79 | 0.73 | -0.69 | | 0.74 | 0.83 | 0.80 | 0.69 | 0.15 | 0.27 | 0.76 | 0.58 | -0.75 | 0.76 | 0.89 | 0.74 | -0.69 | -0.78 | -0.74 |
| ZIMJ680101 | 0.75 | 0.64 | 0.62 | 0.70 | 0.65 | 0.69 | 0.47 | 0.77 | 0.75 | -0.36 | 0.74 | | 0.51 | 0.59 | 0.48 | 0.16 | 0.17 | 0.52 | 0.41 | -0.51 | 0.55 | 0.67 | 0.38 | -0.36 | -0.51 | -0.47 |
| PONP930101 | 0.60 | 0.79 | 0.82 | 0.90 | 0.93 | 0.89 | 0.73 | 0.59 | 0.60 | -0.68 | 0.83 | 0.51 | | 0.68 | 0.67 | 0.29 | 0.46 | 0.85 | 0.56 | -0.72 | 0.74 | 0.78 | 0.91 | -0.68 | -0.91 | -0.75 |
| WILM950101 | 0.66 | 0.54 | 0.64 | 0.65 | 0.73 | 0.68 | 0.73 | 0.75 | 0.66 | -0.61 | 0.80 | 0.59 | 0.68 | | 0.84 | -0.05 | 0.43 | 0.71 | 0.57 | -0.72 | 0.86 | 0.81 | 0.58 | -0.61 | -0.69 | -0.58 |
| WILM950102 | 0.63 | 0.58 | 0.64 | 0.61 | 0.71 | 0.68 | 0.59 | 0.66 | 0.63 | -0.51 | 0.69 | 0.48 | 0.67 | 0.84 | | -0.09 | 0.33 | 0.60 | 0.40 | -0.63 | 0.66 | 0.68 | 0.63 | -0.51 | -0.63 | -0.62 |
| WILM950103 | 0.18 | 0.39 | 0.28 | 0.32 | 0.33 | 0.32 | 0.29 | 0.05 | 0.18 | -0.36 | 0.15 | 0.16 | 0.29 | -0.05 | -0.09 | | 0.15 | 0.25 | 0.17 | -0.41 | 0.16 | 0.19 | 0.49 | -0.36 | -0.49 | -0.35 |
| WILM950104 | 0.36 | 0.21 | 0.29 | 0.38 | 0.49 | 0.37 | 0.21 | 0.36 | 0.36 | -0.19 | 0.27 | 0.17 | 0.46 | 0.43 | 0.33 | 0.15 | | 0.42 | 0.15 | -0.22 | 0.40 | 0.29 | 0.41 | -0.19 | -0.43 | -0.32 |
| JURD980101 | 0.49 | 0.59 | 0.61 | 0.78 | 0.83 | 0.73 | 0.89 | 0.50 | 0.48 | -0.86 | 0.76 | 0.52 | 0.85 | 0.71 | 0.60 | 0.25 | 0.42 | | 0.86 | -0.78 | 0.86 | 0.84 | 0.75 | -0.86 | -0.86 | -0.76 |
| WOLR790101 | 0.35 | 0.34 | 0.37 | 0.56 | 0.61 | 0.50 | 0.91 | 0.38 | 0.35 | -0.88 | 0.54 | 0.41 | 0.56 | 0.57 | 0.40 | 0.17 | 0.15 | 0.86 | | -0.76 | 0.77 | 0.79 | 0.54 | -0.88 | -0.64 | -0.61 |
| KIDA850101 | -0.71 | -0.71 | -0.76 | -0.79 | -0.84 | -0.80 | -0.90 | -0.66 | -0.71 | 0.87 | -0.75 | -0.51 | -0.72 | -0.72 | -0.63 | -0.41 | -0.22 | -0.78 | -0.76 | | -0.87 | -0.90 | -0.82 | 0.87 | 0.86 | 0.79 |
| COWR900101 | 0.68 | 0.56 | 0.66 | 0.75 | 0.80 | 0.73 | 0.86 | 0.69 | 0.68 | -0.78 | 0.76 | 0.55 | 0.74 | 0.86 | 0.66 | 0.16 | 0.40 | 0.86 | 0.77 | -0.87 | | 0.89 | 0.67 | -0.78 | -0.79 | -0.69 |
| BLAS910101 | 0.76 | 0.74 | 0.81 | 0.84 | 0.88 | 0.85 | 0.88 | 0.82 | 0.76 | -0.86 | 0.89 | 0.67 | 0.78 | 0.81 | 0.68 | 0.19 | 0.29 | 0.84 | 0.79 | -0.90 | 0.89 | | 0.75 | -0.86 | -0.79 | -0.83 |
| CASG920101 | 0.59 | 0.79 | 0.80 | 0.83 | 0.90 | 0.86 | 0.73 | 0.52 | 0.59 | -0.73 | 0.74 | 0.38 | 0.91 | 0.58 | 0.63 | 0.49 | 0.41 | 0.75 | 0.54 | -0.82 | 0.67 | 0.75 | | -0.73 | -0.92 | -0.79 |
| ENGD860101 | -0.42 | -0.55 | -0.57 | -0.65 | -0.71 | -0.64 | -0.94 | -0.46 | -0.42 | 1.00 | -0.69 | -0.36 | -0.68 | -0.61 | -0.51 | -0.36 | -0.19 | -0.86 | -0.88 | 0.87 | -0.78 | -0.86 | -0.73 | | 0.76 | 0.83 |
| FASG890101 | -0.60 | -0.79 | -0.79 | -0.85 | -0.90 | -0.86 | -0.80 | -0.55 | -0.60 | 0.76 | -0.78 | -0.51 | -0.91 | -0.69 | -0.63 | -0.49 | -0.43 | -0.86 | -0.64 | 0.86 | -0.79 | -0.79 | -0.92 | 0.76 | | 0.80 |
| GEOD900101 | -0.57 | -0.69 | -0.74 | -0.73 | -0.80 | -0.77 | -0.70 | -0.62 | -0.57 | 0.83 | -0.74 | -0.47 | -0.75 | -0.58 | -0.62 | -0.35 | -0.32 | -0.76 | -0.61 | 0.79 | -0.69 | -0.83 | -0.79 | 0.83 | 0.80 | |

**Figure 3.4 Low consensus in hydrophobicity measures.**
The correlation values for pairs of hydrophobicity from AA index studies are color coded, where green corresponds to more highly positively corrrelated and red represents negative correlation.

In this context, we decided to use the molecular descriptor logP to represent hydrophobicity. LogP measures a subtly different, related property of lipophilicity, which is essentially hydrophobicity with the added consideration of polarity (van de

Waterbeemd 1994). Because lipophilicity is important to understanding the delivery of chemicals through cell membranes (Gombar 1999, Benefanti 2003), it has received concentrated attention and study from the pharmaceutical and pesticide industries, resulting in the careful design and continuous improvement of prediction algorithms. LogP is a specific measure of lipophilicity; the logarithm partition coefficient, which represents the ratio of a compound's concentration in organic versus aqueous-phase solvents of a two compartment system (i.e. a the measure of the molecule's relative solubility in each of the two solvents).

### *3.4c Measuring amino acid charge*

Like hydrophobicity, the electrostatic properties of a given molecule are difficult to quantify. The standard measure, pKa, measures charge – but is highly dependent on the surrounding medium.

We chose to measure the electrostatic interactions of a compound using isolectric point (pI). Whereas other measures of charge are highly dependent on the pH at which the measurement is taken, pI records the pH at which the concentration of the anionic and cationic forms of an amino acid are equal. Experimentally, pI is determined using a titration curve. However it can also be derived theoretically by calculating the pKa (or dissociation constant) values for the ionized states of the amino acid that exist one positive and one negative charge away from the neutral state of the amino acid (Lu 2006). Unfortunately, there are few experimentally determined values for amino acids with which to verify theoretical predictions.

## 3.5 Using chemistry space to investigate the genetic code

The framework of chemistry space provides a powerful tool for looking at the idea of 'early' amino acids (those that were prebiotically plausible and therefore likely to have been the first members of a limited amino acid alphabet: Chapter 2). With this approach we can ask whether any simple patterns unify the arguments made by Weber and Miller for the incorporation of amino acids into genetic coding. In particular, we have noted that a general interpretation of the adaptive hypothesis would expect that the "late" amino acids were selected to augment the protein-building potential of the smaller, "early" alphabet. Indeed, previous researchers have made the qualitative assertions that "*The natural repertoire of 20 amino acids presumably reflects the combined requirements of providing a diversity of chemical functionalities, and providing enough structural diversity that sequences are likely to define unique three-dimensional shapes*" (Hinds, 1996) and that "*The driving force …[in the growth of the amino acid alphabet]… is the possibility to produce fitter proteins when the repertoire of amino acids is enlarged*" (Weberndorfer, 2003). We might therefore expect to detect some sort of expansion of chemistry space associated with the addition of "late" amino acids.

Whereas Weber and Miller considered each amino acid as an individual molecule worthy of a unique story, we are able to view them collectively as points defined by the dimensions of size (as measued by Van der Waals volume) hydrophobicity (logP), and charge (pI). Figure 3.5 therefore shows a simple visualization of the "early" and "late" amino acids (where we followed the "early" designations of Trifonov, as described in Chapter 2) in two dimensional chemistry space using each of the three possible pairings of these physicochemical properties. Although these visualizations lack the complexity of argument presented by Weber and Miller, their simplicity is as much a strength as a weakness under the view that good science simplifies diverse observations into unifying principles. From these plots it appears clear that the "late" amino acids grew the chemistry space of the "earlies" in terms of expanding to greater extremes from an original cluster. As stated above, however, the principle advantage of a quantitative chemistry space is to transform qualitative observations into testable, quantitative hypotheses.

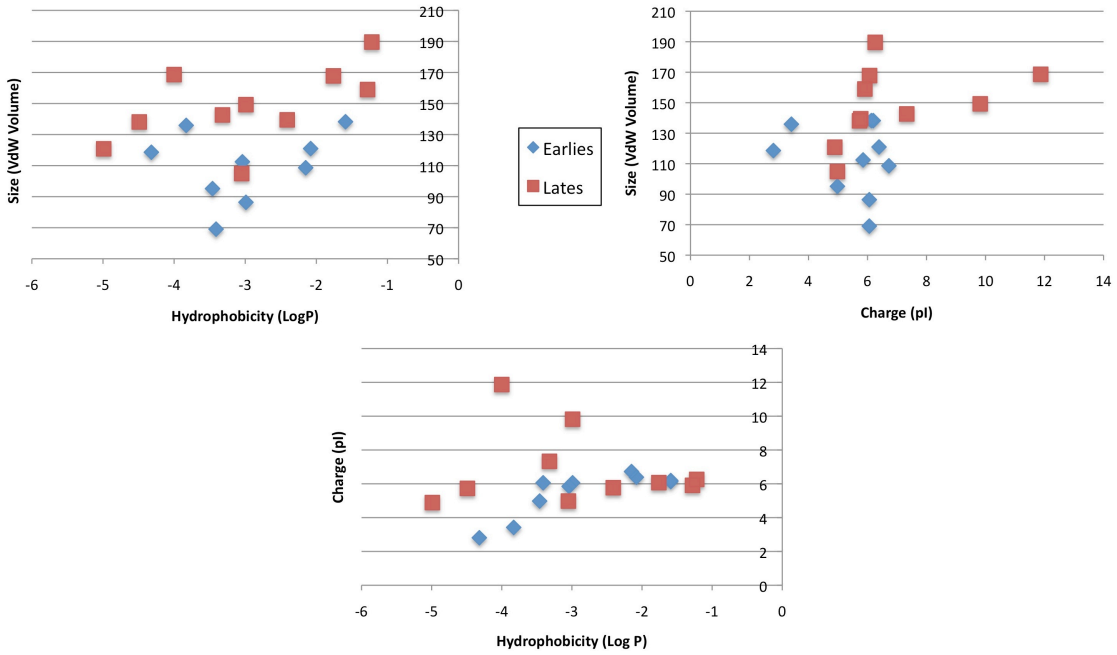



**Figure 3.5 A simple visualization of amino acid chemistry space.**
The genetically encoded amino acids are divided into the "early" amino acids (those which are believed to be prebiotically plausible) and the "late" amino acids, which are generally considered to be 'inventions' of biosynthesis.

If the "late" amino acids truly represent an adaptive expansion of chemistry space, then they should lie further from the "early" cluster than would be expected by chance. A simple to measure this is to take the 20 amino acids and ask: if we were to select 10 amino acids at random and measure their distance from the other 10, then how often would this distance be smaller than the observed difference between "earlies" and "lates"?

In order to test this, we first found the mean of the cluster of "early" amino acids and then measured the distance between this mean and (a) all "earlies", and (b) all "lates". We calculated the ratio of these two distances to give a simple measure of dispersion between the "earlies" and "lates" (see figure 3.6). We then replicated this measure, randomly designating 10 of the twenty amino acids as "early" and 10 as "late." We repeated the process to produce a large sample of random distances that allowed us to record how often the "late" amino acids show greater expansion from the "earlies" than would be expected by chance. By comparing the ratio of the distances to that of random

designations of "early" and "late," we circumvented the problem that a subset of points will often be nearer to its own mean than the other points in the set.

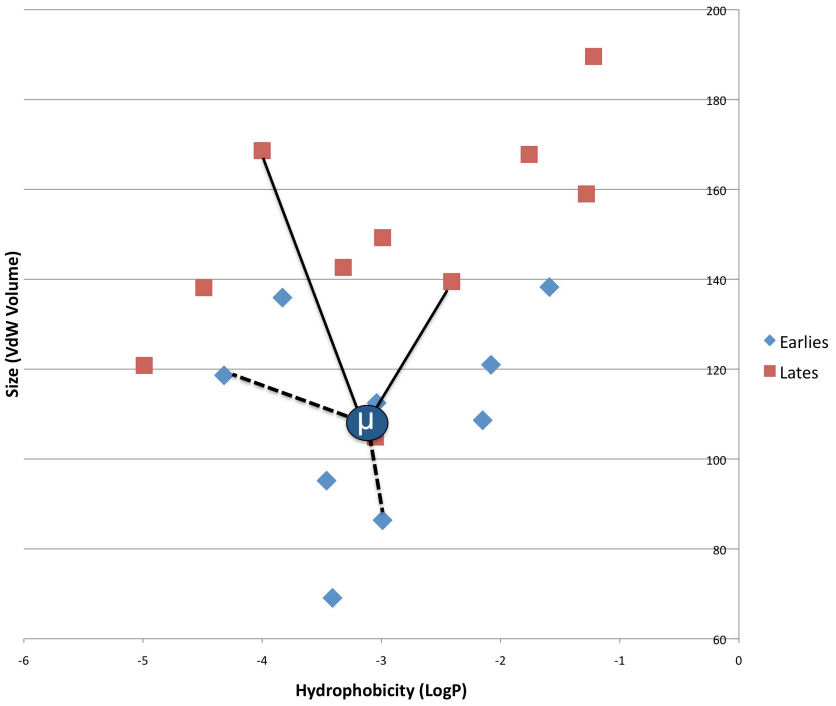
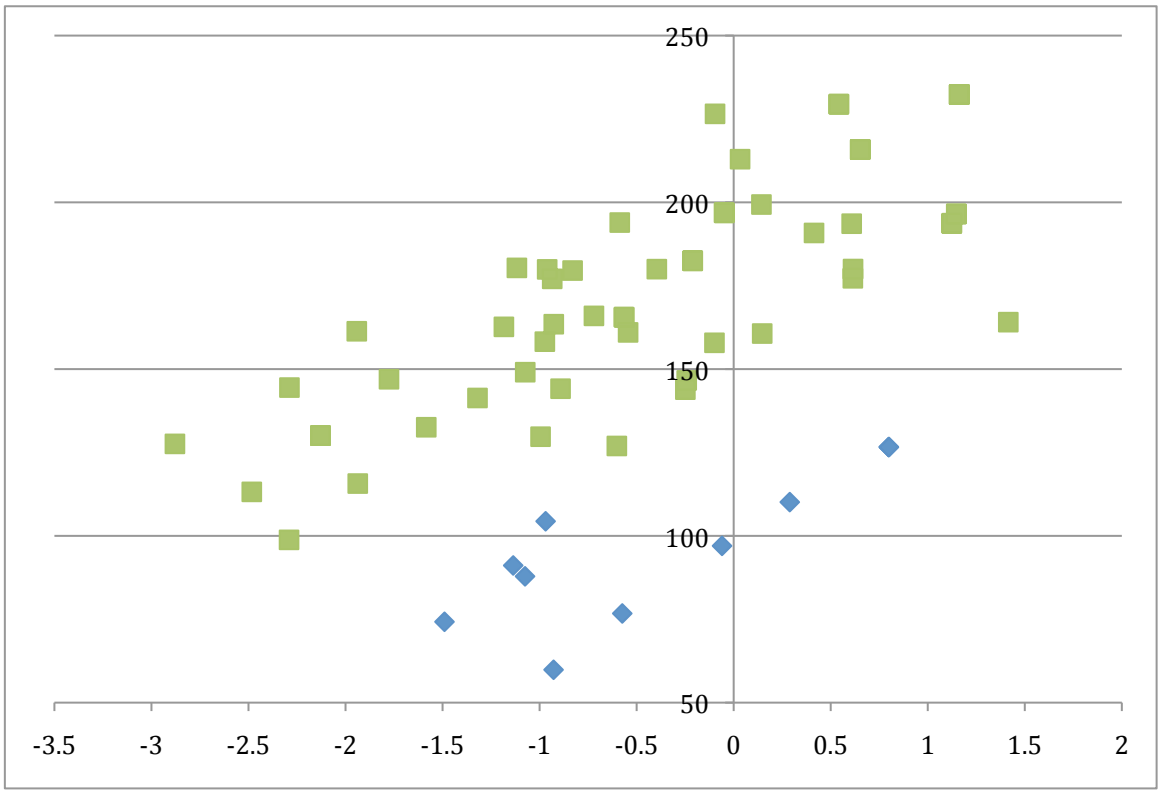



**Figure 3.6 Calculating the expansion of the "late" amino acids from the cluster of "earlies."** This measurement was performed by finding the two-dimensional mean of the "earlies" and comparing the average distance of all "early" amino acids to this mean (represented by the dashed lines) with the average distance from the mean to all "lates" (represented by the solid lines).

Our analysis revealed that the "late" amino acids demonstrate a greater expansion from the "earlies" (as we have chosen to define them) than random assignments of "early" and "late" 99.6%, 99.6%, and 86.7% of the time for the dimensions of size/hydrophobicity, size/charge, and charge/hydrophobicity respectively. These results can be interpreted as broad support of the adaptive hypothesis that the "late" amino acids were adaptively advantageous because they augmented the chemistry space of the "early" amino acids. Importantly, however, the majority of the significance in the first two values appears to be contributed by size. Put simply, the results suggest that the "late" amino acids were bigger. This in itself is not a surprising result, and is explainable in other ways. For example, the distribution of prebiotic amino acids found in meteorites and atmospheric simulations is skewed towards those that are small and easily synthesized by non-enzymatic processes. This alone could potentially account for the fact

that “late” amino acids are, on the whole, larger. However, another simple visualization illustrates that there is a greater effect at work.

Because amino acids in proteins rarely act in isolation, and indeed our concept from the beginning has been that evolution would have selected them as building-blocks for proteins, we expanded our visualization to consider the chemistry space of dimers. Without purchasing chemoinformatics software with which to estimate molecular descriptors of our own choosing, we were limited to such data as already exists. This entailed elimination of the property amino acid charge from our analysis. It also required that we adjust our choice of molecular descriptors to Advanced Chemistry Development Labs (ACD) molar volume for size and ACD LogP for hydrophobicity. Both of these descriptors correlate highly with the descriptors we used in the previous analysis (data not shown).



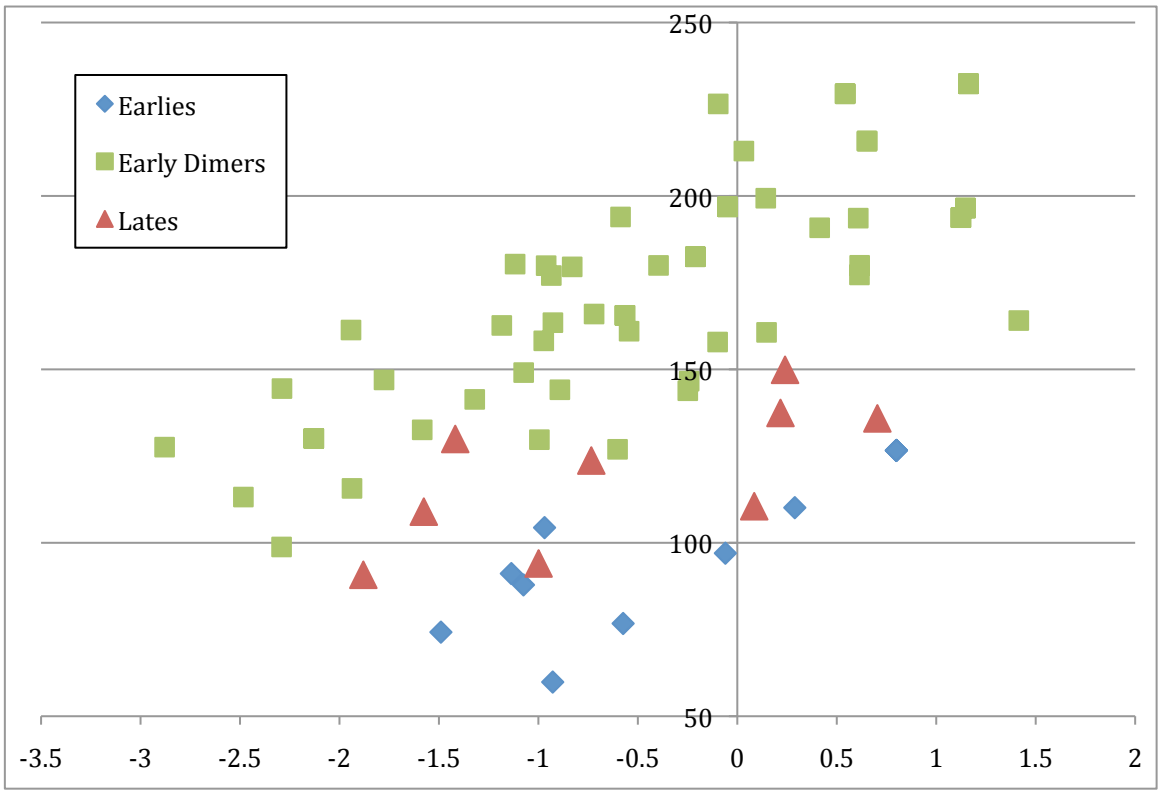



**Figure 3.7 The chemistry space of the early and late amino acids.**
Panel A displays the chemistry space occupied by the early amino acids and the dimers that can be made using only these amino acids, and panel B shows the chemistry space occupied by the late amino acids.

However, even with these limitations, a simple, two-dimensional plot of chemistry space shows clear evidence to support the adaptive hypothesis. The late amino acids are situated within this chemistry space so as to close an otherwise unpopulated gap between early amino acids and the dimers they can form. Without the late amino acids, no single amino acid (or pair of amino acids) could provide the combination of size and hydrophobicity that the "late" amino acids provide: these regions of chemistry space would have been unavailable to early proteins. One might expect the late amino acids to necessarily exceed the size of the early amino acids (after all, they certainly couldn't be smaller than glycine) yet not be as large as an early-early dimer, making them inevitably fall between the early amino acids and their dimers. However, careful inspection reveals that nearly all of the late amino acids are in fact of similar size to at least one early-early dimer: it is the additional factor of hydrophobicity that imbues the late amino acids with a unique role in chemistry space.

This simple intuitive plot, like Dobson's analysis (Figure 3.1), uses simple chemistry space to explain a complex phenomenon. Whereas Dobson's plot demonstrated the unexplored possibility space of bioactive molecules for drug research and discovery, our plot uses the concept of chemistry space to make adaptive, evolutionary sense of the concepts of early and late amino acids.

## 3.6 Summary

In this chapter we have shown how the concept of chemistry space can be used to bridge the gap between the concepts of an adaptive code discussed in chapter 1 and the temporal division of the amino acids described in chapter 2. A simple visualization demonstrates the ease with which profound insights can emerge through the joining of these two concepts. However, in bringing these ideas together we have left open the question of uncoded amino acids. How does everything change in the face of this unexplored, broader pool?

# Chapter 4

## 4.1 Introduction

The previous chapter of this thesis merged three independent ideas: 1) the genetic code is an evolved and evolvable phenomenon; 2) a wide variety of perspectives converge upon the insight that the standard amino acid alphabet may be divided into two sub-sets: "early" amino acids, which were likely available on the prebiotic earth, and "late" amino acids, which were later inventions of metabolism; and 3) the concept of chemistry space can be usefully lifted from pharmaceutical research, where it facilitates quantitative comparisons of biological and non-biological molecules, as an approach to study amino acid alphabet evolution. The chapter ended by demonstrating that a simple plot of amino acid chemistry space offers intuitive, clear support for the general adaptive hypothesis that a simpler, earlier version of the genetic code was augmented with biosynthetically derived "late" amino acids that expanded the protein-building capabilities of the genetic code. Specifically, the "late" amino acids appear to fill a gap that would have existed between the "early" amino acids and the dimers that can be formed by them.

In this chapter I seek to synthesize and extend these ideas by considering the role of natural selection in forming the standard amino acid alphabet given the context of a much larger set of molecular possibilities.

## 4.2 Previous Analysis of Amino Acid Chemistry Space

To date, just two independent analyses have used the concept of chemistry space to investigate the evolution of the standard amino acid alphabet from a broader set of possibilities. One (Zhang 2007 and references therein) used chemoinformatics software to calculate the molecular stability of 8,793 isomers of the 20 standard amino acids. The conclusion was that the amino acids of the standard genetic code are among the most thermodynamically stable options that could possibly have been used as monomeric building blocks for proteins. The significance of this insight is, however, compromised

on two fronts relating to repeatability. First, the database from which amino acids were drawn (the CrossFire Beilstein database) is proprietary and thus the opportunity to verify or build upon this finding is thus limited to researchers who are able to pay for the privilege of accessing its contents. Second, the chemoinformatics software they used made complex thermodynamic stability calculations that are not easily verified (i.e. no other software exists to perform such calculations, and no other approach is clear by which the accuracy of the calculations can be evaluated).

In a second, independent analysis of amino acid chemistry space, Philip and Freeland set out to test evidence that natural selection favored a set of amino acids that exhibit clear, nonrandom properties; put simply, a set of "*especially useful building blocks*" (Philip 2011). This time the researchers emphasized simplicity and repeatability, using easily accessible chemoinformatics prediction software for fundamental molecular descriptors, and building from an in-depth analysis of reliability and repeatability for the descriptor values (Lu 2006).

Moving from this concept they faced three specific challenges. First, it was necessary to establish a background of molecular possibilities against which the standard amino acid alphabet could be compared. In order to represent the pool of L-chiral α-amino acids available to early life through prebiotic synthesis, they looked to the Murchison meteorite (widely considered to be an analogue for the chemical inventory of early Earth, as discussed in Chapter 2). In consideration of the concept of biosynthetic expansion, they included an additional fourteen L-α amino acids that occur as intermediates within the canonical biosynthetic pathways by which the "late" amino acids are formed. Next, they needed to quantify the chemistry space of a particular set of amino acids. For this they used the familiar properties detailed above: size, charge, and hydrophobicity, and identified three specific molecular descriptors (Van der Waals Volume, LogP, and Isoelectric Point) that adequately represented each of these concepts (Lu 2006). Given these dimensions of chemistry space, one final challenge remained: how to quantify the 'coverage' of a set of amino acids.

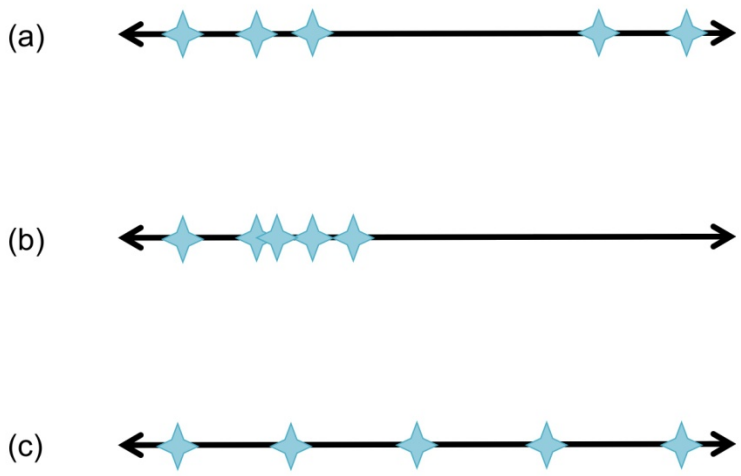



**Figure 4.1 Defining coverage: examples of range and evenness for hypothetical amino acid sets.**
Amino acids are shown as dots on a linear descriptor (e.g. "LogP"). The set defined in (a) displays high range, but poor evenness. In (b), the set is highly even, but with a small range. The set in (c) is optimized for both range and evenness. Adapted from figure in Philip and Freeland (2011).

Philip and Freeland argued that for a set of amino acids to be considered 'useful' they needed (1) to cover a broad range of values within a particular dimension of chemistry space, and (2) to be evenly distributed within that range (Figure 4.1). The rationale of this argument built from the premise that amino acids function as building blocks for proteins that, over evolutionary timescales, are challenged to adapt to constantly shifting demands in terms of the functions they perform and the environments in which they function. To meet these demands, the authors argued that a good set of building blocks should be able to approximate any suite of properties that is needed at any time. In other words, a highly adaptive alphabet would be one that can represent any arbitrary point(s) on the continuous spectra of physicochemical dimensions that define protein activity: size, charge and hydrophobicity. The two properties of an amino acid alphabet's coverage were quantified as *range*, the difference between the maximum and minimum values, and *evenness*, or the variance of the difference between consecutive values (see figure 4.1).

Philip and Freeland then performed three experiments to test this single adaptive hypothesis using slightly different assumptions. In each experiment, the researchers randomly chose one million sets of amino acids drawn from the background possibilities and compared the *range* and *evenness* of size, charge, and hydrophobicity values for each set to those of the genetically encoded amino acids. The logic of this approach was that an amino acid alphabet shaped by natural selection should appear exceptional relative to a large random sample.

The first test compared the set of 20 genetically encoded amino acids to random sets of twenty members chosen from the prebiotic 50 amino acids. This test represented a conservative start-point: if any evidence were to be found for an unusual set of genetically encoded amino acids, then this should be evident by contrasting the "final" contents of the standard genetic code with a random sample of what appears to have been available to an origin of life. The next two tests sought to tighten this reasoning. In the second experiment, sets of eight amino acids were chosen from 50 prebiotic amino acids and compared to the eight genetically encoded amino acids found in the Murchison meteorite. This test represented a putative primordial genetic code that evolved to contain some sub-set of the prebiotically plausible amino acids.

In the third and final test, sets of twenty amino acids were chosen from a background that included the 50 prebiotic amino acids, 14 biosynthetic intermediates, and the additional 12 genetically encoded amino acids and compared with the properties of the entire standard amino acid alphabet. This test was performed to represent the idea that the primordial genetic code evolved through biosynthetic expansion to contain additional amino acids unavailable through prebiotic synthesis. The results of these tests are summarized in figure 4.2. All tests provided strong supportive evidence for the adaptive hypothesis. The chemistry space of the genetically encoded amino acids indeed appears to be highly significant compared to a well-defined, broader set of possibilities.

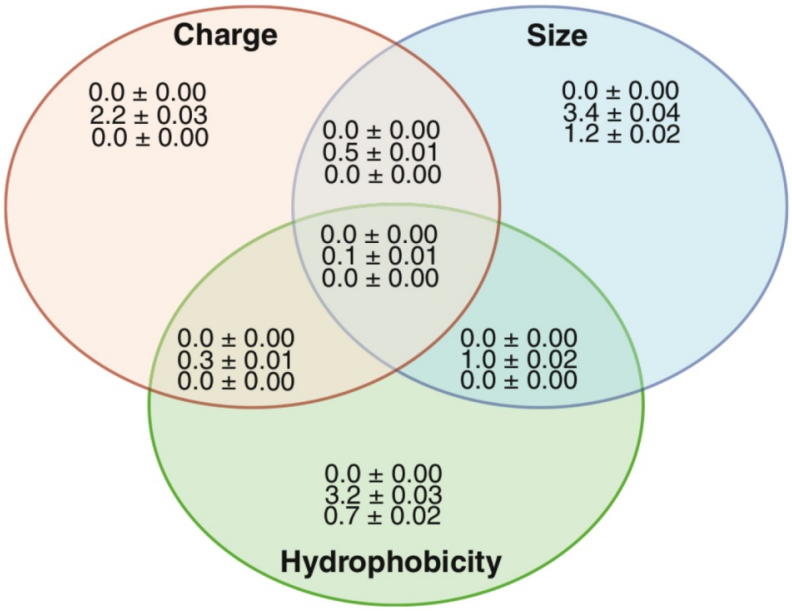



**Figure 4.2 Summary of the results of Philip and Freeland 2011.**
The mean (μ) and 95% confidence interval for the percentage of random alphabets with a coverage (range and evenness) greater than the genetically encoded set for (i) 20 amino acids from a pool of 50 prebiotic candidates (top value); (ii) 8 amino acids (found in the Murchison meteorite) from 50 prebiotic possibilities (middle); and (iii) 20 from 76 (prebiotic + biosynthetic) amino acids (bottom) in each of the three properties of charge (pI), size (van der Waals volume), and hydrophobicity (logP) (from Philip 2011).

If we accept the underlying logic of the tests performed, then the major limitation of this study's findings is that it limited consideration to a pool of at most 76 amino acids, comprising one particular model for prebiotic plausibility (i.e. the Murchison meteorite) and a few biosynthetic possibilities (the 14 intermediates of Wong's hypothesized "molecular fossil" metabolic pathways by which late amino acids were invented). Chapters 1 and 2 of this thesis give good reasons to view both of these pools as highly conservative. Among the many reasons, prebiotic simulation experiments using new instrumentation continue to reveal new abiotically plausible amino acids; hydrothermal vents and other such scenarios offer unknown variation in what is formed; and a modern view of biosynthetic pathways combined with non ribosomal peptide synthesis, post-translational modification, and an exploding view of microbial diversity to suggest we really don't know what evolution is/was capable of producing.

## 4.3 Enlarging our view of the amino acid universe.

A comprehensive exploration of the possibility space of amino acids was recently made possible through the intersection of mathematics, computer programming and organic chemistry (Meringer unpublished). Specifically, research conducted by the research group in which I have been studying has sought to "*explore the universe of chemical structures implied by the description "α-amino acid," with particular emphasis on the size and contents of the set that are relevant to the extant genetically encoded set*" using chemoinformatics. The unique software used in this study, MOLGEN, is a structure generator developed using graph theory (Gugisch 2009). Structure generators employ precise mathematical algorithms to identify every possible chemical structure associated with a given set of inputs (atoms types and numbers, etc.) (Meringer 2010, Gugisch 2007). By representing chemical compounds as graphs, in which nodes represent specific atoms and the edges that connect them represent covalent chemical bonds by which they are linked together, MOLGEN uses rigorous mathematical theory to generate, for the first time, a truly exhaustive set of chemical structures for any given set of atoms.

Much of the challenge in applying the theory (and software) behind structure generators to generating amino acids is to restrict the boundary conditions that define an appropriate set of molecular structures. For example, the authors used the twenty amino acids of the standard genetic code as a guideline to define what functional groups are possible and the maximum size (in terms of carbon atoms) that can occur within a side-chain. Even so, the number of plausible amino acid isomers turned out to be staggering "*far larger than anything that has been suggested to date in the scientific literature*" (Meringer, unpublished). Tryptophan, the largest genetically encoded amino acid, alone had 1.6 trillion isomers, a set that took over 18 days to generate! Indeed, further calculations indicated that a full isomer space for the 20 standard amino acids would produce a library more than an order of magnitude larger than the world's current largest database of chemical structures, a database of drug-like molecules useful for pharmaceutical screening (Blum 2009).

The authors therefore applied further restriction criteria to trim this set of structures into something more manageable. In particular, they steered calculations

towards a conservative interpretation of the most plausible amino acid structures by filtering out molecules that offered any signs of violating considerations of chemical stability or steric strain. In this way, the study led to the definition and construction of a library comprising 4,147 L-chiral α-amino acid structures deemed compatible with chemistry and with what we know of genetic coding.

## 4.4 Re-testing the adaptive properties of standard amino acid alphabet.

The unprecedented data set available from the structure generation software MOLGEN allows us to investigate the coverage of the standard amino acid alphabet compared to a more comprehensive background than ever before.

Figure 4.3 shows the distribution of the 20 genetically encoded amino acids relative to the 4,147 L-chiral α-amino acid structures of this structure generation exercise. Once again, this simple visualization intuitively suggests some interesting phenomena. For example, within the regions of this two-dimensional chemistry space that are populated by plausible amino acids, the largest sub-region that contains few or no representatives within the standard amino acid alphabet coincides with the area that, by comparison with figure 3.7, is in fact populated by dimers of the coded amino acids. This would be consistent with the interpretation that natural selection did not "waste choices" for the coded amino acids by incorporating monomer structures with a size and hydrophobicity implicit to combinations of amino acids already present – an observation that coincides with our previous suggestion that the "late" amino acids close the gap between monomers and dimers. Another interesting aspect to this interpretation is that amino acids tend to become more hydrophillic as they join together with peptide bonds (Figure 3.7). It is thus noteworthy that the largest amino acid monomers are among the most hydrophobic, despite the presence of many less hydrophobic alternatives: once again, the net result is that by the time amino acids are linked together, they populate chemistry space evenly. At this stage, a direct comparison with figure 3.7 (and a quantitative test of this hypothesis) is precluded by a mismatch of the exact molecular descriptors available for dimers and plausible-uncoded amino acids. Put simply, we have only ACD predictors of molar volume and size for dimers; we have only MOLGEN

descriptors available for the large library of uncoded amino acids. Addressing this simple mismatch is therefore discussed as a topic for future research in Chapter 5.

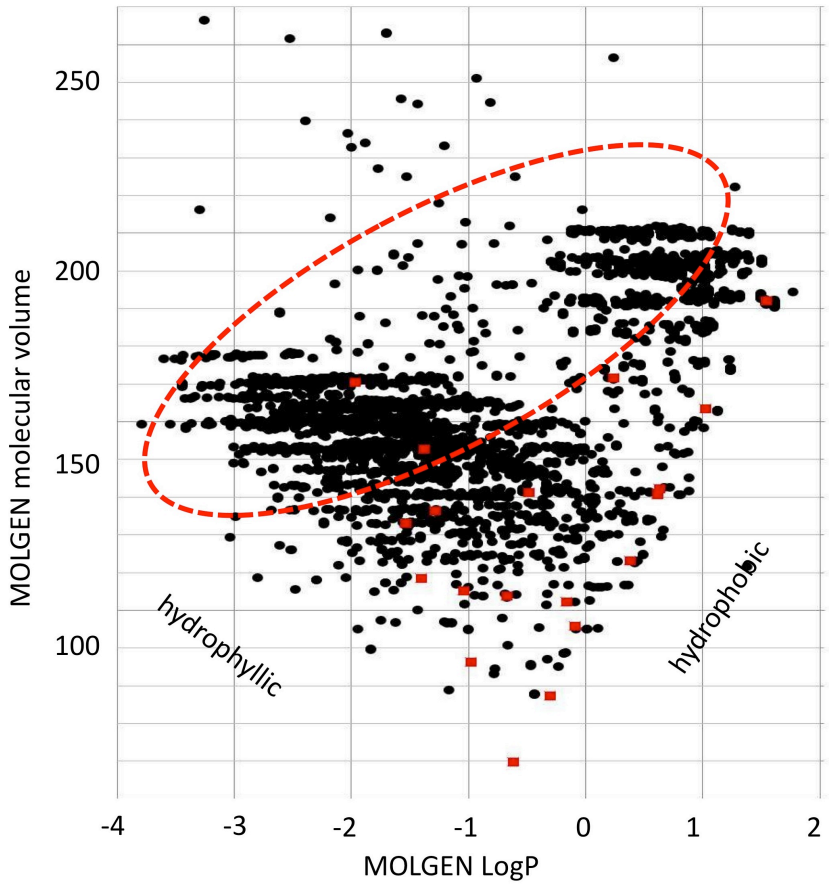



**Figure 4.3 A two-dimensional view of amino acid chemistry space plotted for 4,147 L-chiral α-amino acid structures of plausible relevance to genetic encoding using MOLGEN predictions for LogP (hydrophobicity) and molecular volume (size).**
Black circles indicate plausible structures, red squares indicate the 20 genetically encoded amino acids. Of particular note, the dashed red ellipsis reinforces the observation that genetically encoded amino acids are not found representing sub-regions of chemistry space where dimers of the genetically encoded amino acids occur (as described in Ch. 3, figure 3.7) despite the fact that these regions are heavily populated with chemical possibilities.

A second observation of interest suffers no such limitation. By eye, the data shown in figure 4.3 lend visual credibility to previous findings of unusual "coverage" of chemistry space reported for the genetically encoded amino acids by Freeland and Philip (2011). In other words, the coded amino acids do on the whole appear to cover a relatively broad range, with notable evenness, for combinations of size/hydrophobicity

relative to the much larger pool of structural possibilities. For this observation, we may stand behind our arguments for the usefulness of chemistry space (Ch. 3) by repeating the analysis performed by Philip and Freeland using the extended data set as a definition of alternative amino acid biochemistries that might have been possible through biological evolution.

As per Philip and Freeland's methods, we again sought to measure the coverage (or "usefulness") of the standard amino acid alphabet compared to isomeric alternatives, using the combination of range and evenness described in figure 4.1. This time, however, we chose random sets of amino acids for comparison from a background of 4,147 amino acids rather than 76. We were constrained to the dimensions of size and hydrophobicity due to the limited availability of molecular descriptors.

We performed three experiments in order to analyze the coverage of the genetically encoded amino acids compared to the background of alternatives. In our first experiment, we chose random amino acid alphabets of size twenty from the 4,147 isomers and compared their range and evenness to that of the genetically encoded alphabet. We recorded how often random sets had better coverage in each of the two properties (size and hydrophobicity) individually and in combination. We performed a second experiment in which we selected alphabets of 8 amino acids and compared them to the 8 standard amino acids present in the Murchison meteorite. These two analyses were direct analogues to experiments performed by Philip and Freeland. The third experiment was only subtly different, randomly selecting 10 amino acids and comparing them to the "early" 10 amino acids. Results are summarized in figure 4.4.

In terms of the range of amino acid size (upper left of figure 4.4), there is a strong possibility that the appearance of an unusually "good" range for all 20 amino acids (top value) is an artifact of the fact that the enlarged pool of amino acid structures was deliberately created using the smallest genetically encoded amino acid (Gly) and largest (Trp) as boundaries of the search for plausible structures. However, this consideration does not apply to measures of range in hydrophobicity (lower left of figure 4.3) or to any measure of evenness with which amino acids populate chemistry space (right hand side of figure 4.3). Thus the most highly significant results (range and evenness of any one property, or of both properties combined) are reliable indicators that the genetically

encoded amino acids form an unusual sub-set of chemical structures relative to any random selection of alternatives, whether we consider the standard alphabet as a whole, or theorized, plausible stages of its evolution at which it comprised fewer molecules.

In other words, this analysis demonstrates that the earlier work by Philip and Freeland is not an artifact of any narrow view of prebiotic plausibility or biosynthetic availability, but instead the adaptive qualities of the standard amino acid alphabet are exceptional relative to our best guess at structural possibility space for L-α -amino acids.

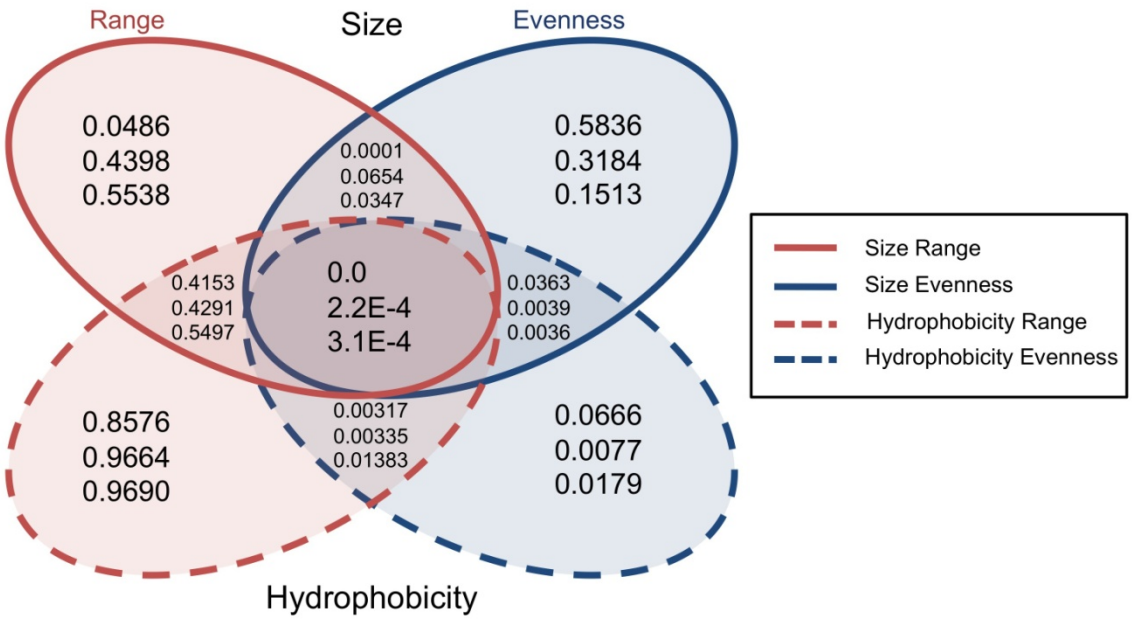



**Figure 4.4 The coverage of the standard amino acid alphabet compared to a comprehensive background of possible isomers.**
In each sub-section of the Venn diagram, the three values shown indicate what proportion of randomly selected sets (from a pool of 4,147 amino acids) exhibit better range, evenness, or a combination for size and hydrophobicity. The top value represents randomly selected alphabets of size 20 compared to the standard 20, the middle value represents sets of size 8 compared to the 8 amino acids found in the Murchison meteorite, and the bottom value shows sets of size 10 compared to the "early" 10. For each, 100,000 random sets were selected.

## 4.5 Summary

This chapter presents novel evidence that the standard amino acid alphabet exhibits exceptional properties when compared with a background of plausible L-α amino acid alternatives that have never before been explored for this purpose.

Remarkably, the same conclusions reached previously by considering the standard amino acid alphabet to a background of at most 76 amino acids match those obtained when comparing the alphabet with a background that is orders of magnitude larger. This adds weight to the adaptive hypothesis for the origin of the standard genetic code, suggesting that natural selection may have played a far greater role than can be inferred from a simple consideration of the pattern by which amino acids are assigned to codons.

The most obvious limitations of the analyses presented here concern the lack of molecular descriptors currently available for the larger set of chemical structures we consider. In particular, we are unable to analyze the data set in the dimension of charge until we are able to obtain commercial chemoinformatics software capable of extensive descriptor prediction. This limitation is particularly interesting when we note that it is precisely when we start to combine dimensions of size and hydrophobicity that the standard amino acid alphabet appears most special (Figure 4.4).

However, the point of this thesis is not to present a comprehensive answer to the question of how and why the standard amino acid alphabet evolved. Rather, it is to demonstrate the power and possibilities of an emerging approach that uses computers to generate plausible alternatives to biology-as-we-know it; generating chemical structures, predicting their properties, and measuring the ways in which life distinguishes itself from randomness. Twenty years ago, computers were being used to demonstrate new evidence that could replace the "Frozen Accident" explanations for the standard genetic code with a logic of natural selection (Haig 1991, Freeland 1998). Ten years ago, tools of bioinformatics were being used to demonstrate from RNA sequence data that a stereochemical basis for the genetic coding left plenty of room for natural selection of an adaptive pattern of codon assignments (Caporaso 2005). This thesis aims to reveal that chemoinformatics and chemistry space combine to offer a new and exciting direction with which to probe further into the origin and evolution of the genetic code.

# Chapter 5

## 5.1 Introduction

The aim of this thesis is to describe and then build upon current scientific understanding regarding the evolution of the standard amino acid "alphabet" with which life on Earth evolved to build genetically encoded proteins. Consistent with the nature of scientific research in general, this body of work leads to many more questions than it answers. The purpose of this final chapter is to highlight some of these directions for future research, with an emphasis on those for which near-term future research may reasonably expect to yield progress. Most clearly, this includes specific suggestions for research within the relatively well-defined domains of biochemistry, bioinformatics and chemoinformatics (section 5.2). However, a significant fraction of the insights drawn in earlier chapters originate within other scientific disciplines, such as geochemistry and planetary science, which have a less well-developed interface with life science. The emerging "meta"-science of astrobiology explicitly seeks to remedy this interdisciplinary divide (Desmarais 2008), and section 5.3 concludes by highlighting insights that seem plausible outcomes of imminent space exploration.

## 5.2 Suggested future directions

The status of research described in preceding chapters suggests at least three research goals that, through a reasonable investment of time and resources, could improve our understanding of amino acid alphabet evolution: (i) a chemoinformatics analysis to understand the physicochemical basis for the emergence of the standard amino acid alphabet; (ii) a bioinformatics analysis of microbial diversity to update evidence for the biosynthetic expansion of the amino acid alphabet; and (iii) empirical and/or bioinformatics research to understand the protein-building implications of a smaller, earlier amino acid alphabet.

### 5.2a Further chemoinformatics analysis of amino acid alphabet properties

The novel research presented in this thesis arises from bringing the concept of chemistry space to bear upon hypotheses for the emergence of the standard amino acid alphabet. In particular, chapters 3 and 4 advance evidence to corroborate and extend an adaptive hypothesis for the emergence of the standard genetic code by noting non-random, plausibly adaptive attributes of the standard amino acid alphabet relative to other possible suites of amino acids. In other words, amino acid chemistry space shows a signature consistent with natural selection for a powerful and versatile set of building blocks.

The major limitation on the results presented is their reliance on pre-existing data for amino acid molecular descriptors. Chapter 3 uses estimates of molecular volume (size) and LogP (hydrophobicity) both derived from the ACD (Advanced Chemistry Development) Labs' PhysChem Suite[2] as these data are provided free of charge on the ChemSpider web resource[3] by a research group within the Royal Society for Chemistry. The resulting view of amino acid chemistry space (Figure 3.7) incorporated for the first time a context of amino acid dimers alongside the individual amino acids of the genetic code. This was possible simply because appropriate molecular descriptor values were already available. For exactly the same reason, in its present state this work fails to connect with previous adaptive analysis of the standard amino acid alphabet (Philip 2011) on two fronts. First, it fails to incorporate the third dimension of amino acid chemistry space (charge) that has previously been linked to adaptive expansion of the genetic code; second, the visualization shown in figure 3.7 uses subtly different molecular descriptors to represent the conceptual properties of size and hydrophobicity than this previous analysis. These mismatches reflect the practical difficulties of obtaining the molecular descriptors used by Philip and Freeland for the broader set of amino acids and dimers that we consider here. Similarly, the analysis presented in Ch. 4 fails to include consideration of amino acid charge because it is limited to the predictions available through the

[2] http://www.acdlabs.com/products/pc_admet/physchem/physchemsuite/
[3] www.chemspider.com

chemoinformatics prediction software, MOLGEN 5[4], available to our research group during this thesis.

This state of affairs reflects the underlying fact that an overwhelming majority of research concerning amino acid molecular descriptors has focused entirely on the amino acids found within the standard genetic code. For example, the AAIndex database[5] contains more than 500 different quantitative measurements for the amino acids of the standard alphabet, and none for any other molecules. It is still new and unusual thinking to extend such measures to incorporate the much larger pool of amino acids relevant to thinking about life's origins and evolutionary/synthetic possibilities.

Perhaps the simplest step forwards from this thesis would therefore be to render a comprehensive matrix of amino acid descriptor values that extends the size, charge, and hydrophobicity values discussed above. This would require little more than the procurement of accurate, reliable chemoinformatics software: too expensive for a graduate budget but well within the scope of a small research grant. This would not only close the gaps between present and previous analyses of amino acid chemistry space, but would provide a robust foundation for further quantitative hypothesis testing of theories for amino acid alphabet evolution. From here, a further useful step would be to build a free, online database, analogous to the AAIndex but dedicated to collecting measurements that systematically represent the larger molecular context of amino acids. One way in which to approach this latter challenge would be to seek collaboration with preexisting chemical databases, such as ChemSpider, to facilitate the study of amino acids as a specific sub-class of molecule. It is perhaps appropriate to note that such an initiative would match the spirit of free, centralized information that has characterized the early development of bioinformatics, catalyzing the rapid growth of this field relative to chemoinformatics.

### *5.2b Bioinformatics analysis of amino acid biosynthesis pathways*

As chapter 2 describes, one of the central ideas for the emergence of the standard amino acid alphabet asserts that a genetic coding originated with a smaller amino acid

[4] http://molgen.de/?src=documents/molgen5.html
[5] http://www.genome.jp/aaindex/

alphabet; those that are reliably produced by prebiotic syntheses. This original genetic code then expanded as early evolution invented new amino acids and selectively incorporated them back into the genetic code.

This idea originated from a biological claim (Wong 1975) that extant pathways of amino acid biosynthesis are "molecular fossils" of this expansion process. That is, prebiotically plausible amino acids lie at the head of modern biosynthetic pathways, and prebiotically implausible amino acids lie at their termini precisely because these pathways are unchanged since evolution first created additional amino acids by these very pathways. Since the launch of this idea, the underlying claim for subdividing the standard amino acid alphabet into those that were present from the beginning ("earlies') versus those that emerged subsequently ("lates") has gathered considerable evidence from a variety of disciplines, including multiple meta-analyses (e.g. Trifonov 2001, Higgs 2009, Cleaves 2010). At the time of writing this thesis, however, the most questionable line of evidence for "early" versus "late" amino acids is the original biological evidence from which the hypothesis was derived. Subsequent decades of research, especially tools of molecular biology, have opened up a view of previously unknown diversity in microbiology and associated metabolic pathways that suggest that a re-evaluation of Wong's biosynthetic pathways is in order.

In this context, another well-defined goal for future research would be to obtain a detailed view of amino acid biosynthesis pathways based on an updated view of metabolic diversity. This analysis could be used to evaluate exactly which (and how many alternative) pathways of amino acid synthesis are plausibly ancient – dating back at least as far as LUCA. The investigative logic is relatively simple and has been well described for those interested in establishing the genetic composition of LUCA (Benner 1989): to be a plausible ancestral pathway, the enzymes involved should be present in at least two (preferably all three) domains of life at a frequency and distribution that removes any serious doubts of being an artifact of "recent" lateral transfer. Such analysis would be relatively straightforward given the existence of databases such as KEGG (the Kyoto Encyclopoaedia of Genes and Genomes, Kanehisa 2004), BRENDA (BRaunschweig ENzyme DAtabase, Schomburg 2002) and MetaCyc (Caspi 2008),

which offer carefully curated views of genetic, enzymatic and metabolic pathway data respectively.

A detailed study of biosynthetic pathways would not only verify the validity of Wong's claims, but would also paint a clearer picture of the biosynthetic intermediates that should be considered when studying which amino acids evolution might plausibly have 'chosen' to include in its genetically encoded alphabet. Thus far, the only study to estimate this set of molecules identified a total of 14 intermediates in the original pathways highlighted by Wong (Philip 2011). Other studies have estimated that thousands of additional amino acids occur within microbial metabolism (Uy 1977), an unknown number of which might date back to LUCA and beyond (see Freeland 2009 for a review).

### *5.2c Protein-building implications of alternative amino acid alphabets*

Given the pre-eminence within current thinking that the amino acid alphabet started with fewer than the standard 20 amino acids, a third general frontier of inquiry would be to ask what implications a smaller earlier amino acid alphabet would carry for the earliest protein fold universe. Although the "protein folding problem" remains a very real challenge for 21$^{st}$ century biologists (Moult 2011), one of the most exciting insights into the complex connections between amino acid sequences and protein structures is that the indefinitely large suite of protein structures found in today's biosphere may be meaningfully grouped into a finite number of underlying protein "folds" or "domains" (Kolodny 2013). It also appears that microbial evolution, which comprises an overwhelming majority of the history of life on Earth, was primarily responsible for populating this universe of different folds – evolution within multicellular lineages seems to have focused more on the assembly and shuffling of these folds into diverse "multidomain" complexes (Levitt 2009).

At present, it is almost entirely unknown which of the architectures that comprise this "protein fold universe" would have been available to living systems that genetically encoded at most half of the standard amino acid alphabet. Would such an alphabet have permitted the construction of some specific subset of folds, or would it have enabled some crude approximation of all (or at least most) classes of folds? More fundamentally,

it is unknown whether a simpler, smaller amino acid alphabet would have made it easier or more difficult for random, non-synonymous mutations to explore useful protein structures and "discover" new folds. This question is becoming increasingly tractable through recent advances in both bioinformatics and empirical science.

Computational investigations have benefitted from software packages such as ROSETTA (Leaver-Fay 2011), which are making significant progress in providing meaningful *ab initio* folding predictions from user-defined amino acid sequences. These packages are able to make sequence-to-structure predictions independent of alignments with known protein sequences and structures. This provides an emerging opportunity to probe sequence/structure mapping with computer-generated sequences limited to "early" amino acids. One of the most accessible ways to start refining specific hypotheses would be to measure how a library of structures comprising theoretical sequences made of only "earlies" compares with equivalent libraries comprising sequences that utilize the full amino acid alphabet in terms of the percentage of folds or fold-like sequences that each contains.

In the world of empirical science, a number of different research groups are pioneering the artificial selection of protein structures and functions independent of the standard genetic code (e.g. Golynskiy 2013). One group has pioneered an approach to explore how structure and function occur within combinatorial libraries that were not "*evolutionarily selected to perform any particular type of activity*" (Patel 2012). In particular, this work uses "*binary patterning of polar and non-polar amino acids*" such that there is, in principle, no reliance on the standard amino acid alphabet. This technology could be used to explore empirically the ease with which a reduced amino acid alphabet can "find" useful structures and functions, and perhaps even the range of such biologically meaningful molecules that are accessible.

Beyond the applications of existing tools and technologies, there is yet another unexplored branch of this concept. Could a different alphabet of amino acids lead to entirely new folds? There currently exists almost no scientific literature to indicate whether the range of protein folds we experience on Earth is contingent upon the amino acids from which it is constructed, or whether physical considerations force variations in the amino acid alphabet to converge upon particularly stable and inevitable subsets of

secondary and tertiary structures. The one entirely theoretical claim that explicitly addresses this question has argued for the latter, counter-intuitive suggestion but has yet to find any empirical support (Denton 2002).

The most promising approach for this ambitious research goal might presently be to capitalize on the remarkable success of synthetic biology in engineering artificial genetic codes. To date, this technology has allowed more than 70 "unnatural" amino acids to be absorbed into the genetic code (Liu 2010). As such methods continue to mature and expand (Johnson 2011), coupled with the combinatorial techniques described above, it seems only a matter of time before explorations of theoretical protein fold universes become an empirical reality. Indeed, any progress towards answering such questions would not only inform our sense of possible alternative biological evolution but could offer important insights to genetic engineering, where substantial financial gain awaits "unnatural" (and therefore patentable) protein products.

## 5.3 Astrobiological frontiers for understanding the amino acid alphabet

Scientific investigations surrounding the origin of life necessarily expand beyond biology. Indeed, the entire question of life's origins requires an understanding of how replicating, evolving systems emerged from the non-living environment. The inert materials from which life originated remain directly relevant to understanding the "choices" early biological evolution made in order to fashion subsequent metabolism. As the early chapters of this thesis demonstrate, considerable insight into the origin and evolution of the standard amino acid alphabet has come from beyond the traditional boundaries of life-science. In particular, attempts within chemistry to simulate prebiotic conditions must draw from planetary science (including geoscience and related sub-fields of planetary astronomy) to inform what range of amino acids would have been available at the dawn of protein synthesis.

Scientific literature from chemistry and the life-sciences still routinely cite Miller's spark-tube experiments as a foundational insight into the origin of important biomolecules. However, planetary science has significantly changed our understanding of early Earth, particularly its atmosphere, since the time of Miller's experiments. In 1953, Miller assumed an atmosphere comprising a "strongly reducing" gas mixture of $H_2$,

$H_2O$, $CH_4$, and $NH_3$ in accordance with what at the time was commonly believed to be representative of the early Earth (Miller 1953, Urey 1952). Subsequent geoscience (e.g. Kasting 1993) has reached relatively broad consensus that photochemistry would have led to large-scale decomposition of $CH_4$ and $NH_3$. Early Earth's atmosphere would instead have been a "weakly reducing" mixture, dominated by $N_2$, $CO_2$ and CO (with smaller quantities of $H_2S$, $CH_4$, and $H_2$). The relevance of this shifting view is that variations of the spark-tube experiments, which assume a weakly reducing atmosphere, can significantly decrease in the diversity and abundance of amino acid synthesis reactions (Schlesinger 1983), most likely because of the lowered synthesis of hydrogen cyanide as an important intermediate (Ferris 1978).

Controversy and debate continue to surround this topic. Heterogeneous surface conditions could well have included micro-environments, such as volcanic plumes, with highly reducing conditions (Urey 1952, Johnson 2008, Tian 2005, Walker 1985) and synthesis of amino acids may not be as dependent on reducing conditions as is generally thought (Cleaves 2008, Plankensteiner 2006). This debate emphasizes that research into the concept of "early" amino acids (those that were, by virtue of prebiotic plausibility, available to the earliest living systems) must synthesize a multi-disciplinary array of insights to reach sensible conclusions. It also highlights a central challenge for astrobiology: how can we get past the fact that we presently know of only one instance of life (the "N=1" limitation) in order to expand scientific understanding of how and with what ease life originates within an abiotic cosmos? To address this limitation, I present two suggestions for further research that focus upon imminent progress in space exploration. Each offers specific opportunities to test current scientific understanding of the origin and evolution of the amino acid alphabet by looking beyond Earth.

### *5.3a Amino acids on Comets*

Comets are often described as the most pristine examples of material from which the solar system formed, in that they condensed from molecular clouds during solar system formation without ever experiencing temperatures and/or pressures to alter their

chemistry or physical attributes[6]. Of particular relevance to the topic of this thesis, extensive astronomical analyses indicate that comets primarily consist of frozen water, ammonia, methane, nitrogen and carbon dioxide. In other words, comets not only contain, but largely *comprise*, the ingredients used in Miller's original spark tube experiments (see Chapter 2). Indeed, part of Miller's work included elegant tests to demonstrate that the pathway of amino acid synthesis in spark tubes was a reaction long known to chemistry as the "Strecker Synthesis" (Miller 1959) which requires nothing more than ammonia, methane and water (see figure 5.1). Therefore, if our conception of early amino acids is correct, we might expect to find early amino acids formed within comets. There is however one major caveat to this statement: the Strecker Synthesis requires *liquid* water (and a source of energy).

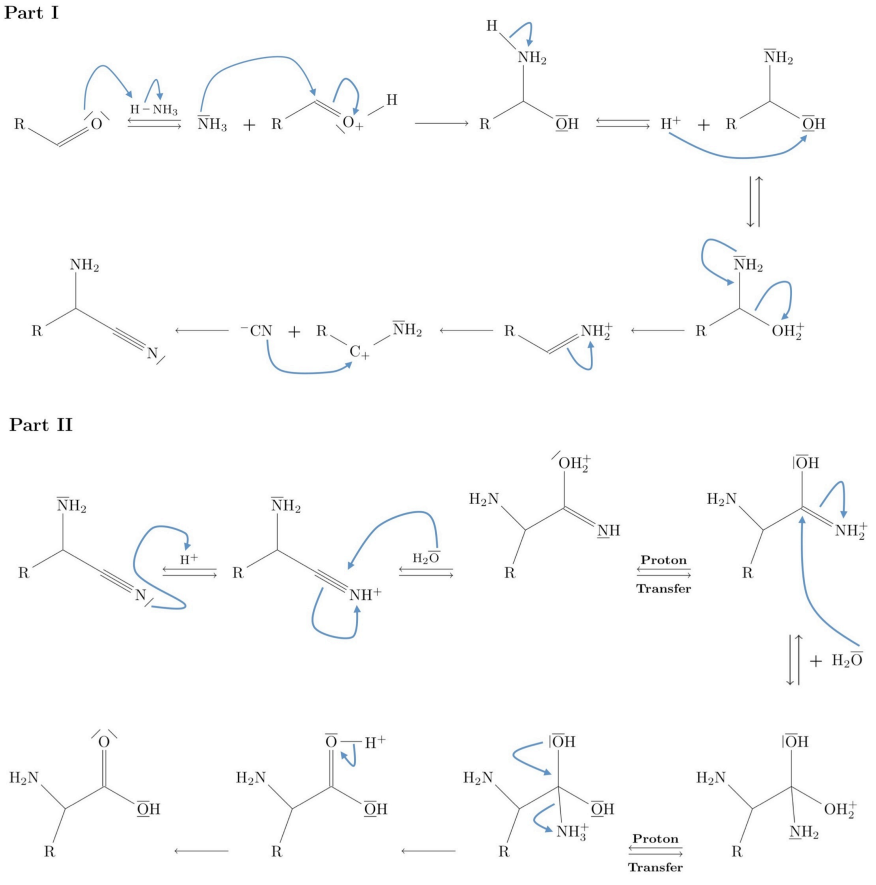



**Figure 5.1 The mechanism of Strecker Synthesis of Amino acids.**

[6] "What are comets and asteroids?" http://neo.jpl.nasa.gov/faq/#ast

Comets are "pristine" in part because they formed in the outer solar system where low temperatures ensure that water is always present in the form of ice. Of course, comets are sometimes sent hurtling towards the inner solar system. In the process of approaching the Sun they typically "come to life" in the sense that rising temperatures cause ices to melt off gas and dust that are blown away by solar winds to form the comet's vast coma that we are occasionally fortunate enough to see lighting up the night sky. This sounds hopeful for amino acid synthesis, but even here the low pressures of space are such that water ice sublimes directly into gas. Perhaps unsurprisingly, then, empirical work has largely failed to detect amino acids associated with comets. Until recently, comet composition data was limited to the findings of ground-based observation via spectroscopy (amino acid chemical structures are at the upper end of complexity for signatures that existing spectroscopic analysis can detect via remote observation). In 2006, The Stardust mission was able to return samples from the Wild-2 comet collected during a 'fly-by' of the coma (the "envelope" surrounding the solid, central component or "nucleus" of the comet) (Flynn 2006). The coma dust was collected in an aerogel sample, and it was initially believed that glycine had been detected above control levels in the dust. However, isotopic analyses remain uncertain whether the detected amino acid was terrestrial in origin (Sanford 2006, Elsila 2009), and this ambiguous signal for glycine remains the only suggestion of amino acids within comets to date. That observation is, however, curiously analogous to the results of the first Miller atmospheric simulation experiment (Miller 1953), which many forget only definitively identified a single amino acid (also glycine). The extensive literature on amino acids that have been identified in similar experiments since that time (discussed at length in chapter 2) reveal that initial detection was severely limited by instrumentation. Given that the Stardust mission was the first experiment of its kind, it is not unlikely that the results are also limited, particularly when one factors in the decades that it takes for scientific instrumentation to gain "clearance" for use on spacecraft: we have good reason to be cautious of interpreting current data on the organic chemical inventory of comets.

Interestingly, recent research in theoretical astronomy is starting to question the notion that cometary water always sublimates directly from ice to gas. Theoretical calculations suggest that liquid water could potentially exist within comets, particularly

as a result of internal heating from radioactive decay of aluminum$^{26}$ (Prialnik 2008). If that is true, it would be surprising if amino acids were *not* synthesized under such conditions.

Fortunately, the upcoming European Space Agency "Rosetta" mission will provide us with an imminent chance to explore further the potential for organic synthesis in these extraterrestrial bodies by actually landing on a comet's surface. Launched in 2004, the Rosetta spacecraft will, in 2014, deploy the Philae lander to the nucleus of the comet 67P/Churyumov–Gerasimenko. This lander is equipped with instruments that should provide considerable new insight, including the COSAC (COmetary SAmpling and Composition experiment), which will "*detect and identify complex organic molecules from their elemental and molecular composition*." (Goesmann 2005). On the basis of the information presented in chapter 2, we may go so far as to predict a specific inventory of "early" amino acids that would test ideas for amino acid alphabet evolution.

### *5.3b Amino acids and the search for life on Mars*

As our nearest planetary neighbor, Mars has long been the subject of speculations and investigations relating to extraterrestrial life. Aside from exuberant claims that have failed subsequent scientific investigation (including Percival Lowell's early observations of seasonal vegetation blooms which are now interpreted as giant dust storms, and the ALH84001 meteorite which was subsequently identified as originating from an underground magma chamber) it is noteworthy how little confidence we can have that the red planet hosts no living organisms. To date, only the 1976 Viking mission explicitly focused on testing for extant life, and even here ambiguous results for one of the three experiments carried out have left a significant minority skeptical of a non-biological interpretation (Klein 1978, Navarro-Gonzales 2006). However, any focus on extant Martian life misdirects attention from the true value of Mars to understanding life's origin on Earth.

In addition to its proximity to Earth, the red planet is a relatively close geophysical analog to our planet. Similarities include bulk planetary composition (Taylor 2006), the prevalence of water (including liquid water) (eg McKay 1991), and even some suggestion of early plate tectonic activity (currently the only known example of

extraterrestrial plate tectonics) (Connerney 1999). Resemblance to Earth is particularly strong during Mars' warm, wet early geological history 4.1-3.7 billion years ago (the "Noachian" period), which coincides with current estimates for the time at which life on Earth originated - the transition from the Hadean to Archaean periods). Thus the opening paragraph to a 1995 NASA report entitled "An Exobiological Strategy for Mars Exploration" explains *"…[owing to similarities between Mars and Earth] A determination of how far Mars proceeded along a path towards life would be of fundamental significance by greatly improving our definition of the window of opportunity within which life could originate. It is important to note this remains true whether or not evidence is found for present or former life on Mars*" (Kerridge 1995).

The search for evidence of organics on Mars is especially promising since the same processes that have erased traces of the earliest molecules on Earth through recycling (namely scavenging by living systems and extensive plate-tectonic activity) have not occurred on Mars. It is therefore likely that the Martian surface has preserved indications of this early chemistry far better than anything we will ever find on Earth (Albarede 2009).

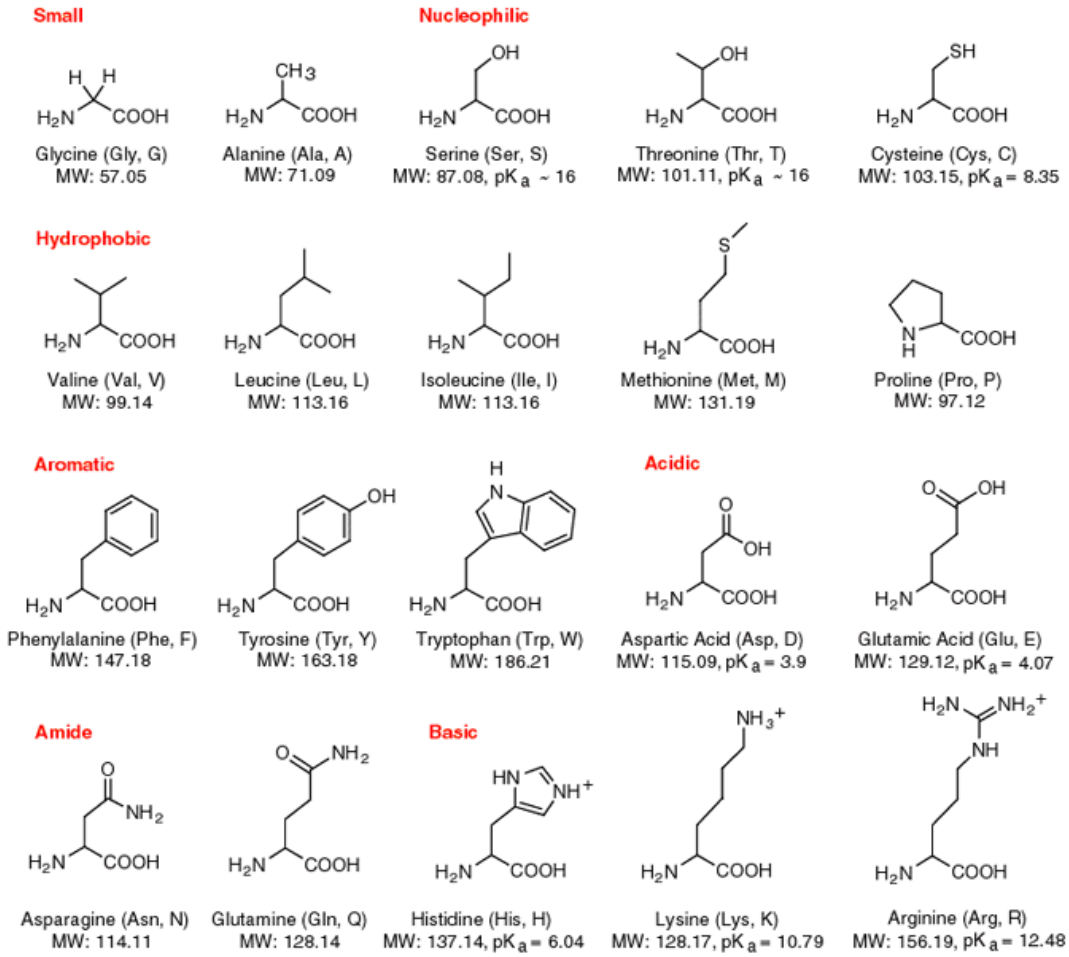
A notable source of organic material on Mars is meteorites. Organics present on these impactors would have a high probability of surviving Martian delivery due to the thin atmosphere surrounding Mars compared to Earth (ten Kate 2010). Additionally, the planet's lower gravitational attraction would result in a reduced impact velocity. Since meteorites have been proven capable of delivering organics to Earth (eg. Kvenvolden 1970), it is widely viewed that they could have been be a significant source of organic compounds on both an early Earth and early Mars (Chyba 1992, ten Kate 2010). Given the probable presence of water at Mars' surface, organics are also expected to be present due to two methods of planetary synthesis. The first, weathering, describes the interaction of liquid water on and below the surface of Mars with igneous rock through flows and melting events (Kerridge 1995). The second, a consequence of liquid water paired with Mars' geological activity, is hydrothermal organic synthesis. This theory is supported by evidence of aqueous alteration in SNC meteorites (so named for their subclasses; Shergottite, Nakhlite, and Chassigny) believed to originate in hydrothermal systems and involving the exchange of water between the surface and subsurface.

Fortunately, any ambiguity in our understanding of Martian surface chemistry is likely to be at least partially addressed presently thanks to the Mars Science Laboratory (MSL) otherwise known as the Curiosity rover, which is currently exploring Mars' Gayle crater. One of the suites of instruments aboard the rover, Sample Analysis at Mars (or SAM) is specifically designed to "*address the present and past habitability of Mars by exploring molecular and elemental chemistry relevant to life*"[7]. One of the "relevant" classes of molecules that SAM will seek to identify is amino acids.

As we receive data from SAM and the Philae lander about the presence of organics on Mars and comets respectively, it seems likely that we will soon witness progress in our understanding of what amino acids are truly prebiotically plausible and in what abundance they were available in the early solar system. Improved knowledge of which amino acids can truly be classified as "early" would also help to constrain the list of amino acids that should be considered "lates." If specific amino acids can be concluded to necessarily be products of life, then those amino acids could potentially used as biosignatures in future astrobiological exploration of extraterrestrial environments. "Late" amino acids whose production we cannot understand in the absence of life, if identified, could be a signal that some kind of microbial metabolism is occurring.

[7] Sample Analysis at Mars (SAM) (**msl**-scicorner.jpl.nasa.gov/Instruments/**SAM**/)

# Appendices




## Appendix A

The structures of the twenty genetically encoded amino acids and their one and three letter abbreviations.

## References


Albarede, Francis. "Volatile accretion history of the terrestrial planets and dynamic implications." *Nature* 461.7268 (2009): 1227-1233.

Alff-Steinberger, C. "The genetic code and error transmission." *Proceedings of the National Academy of Sciences* 64.2 (1969): 584-591.

Amend, J. P., and E. L. Shock. "Energetics of amino acid synthesis in hydrothermal ecosystems." *Science* 281.5383 (1998): 1659-1662.

Ardell, David H. "On error minimization in a sequential origin of the standard genetic code." *Journal of molecular evolution* 47.1 (1998): 1-13.

Bakermans, Corien. "Limits for microbial life at subzero temperatures." *Psychrophiles: from Biodiversity to Biotechnology* (2008): 17-28.

Barker, Andy, Kettle, JG., Nowak, T., Pease, JE. "Expanding medicinal chemistry space." *Drug discovery today* (2012).

Bashford, J. D., I. Tsohantjis, and P. D. Jarvis. "Codon and nucleotide assignments in a supersymmetric model of the genetic code." *Physics Letters A* 233.4 (1997): 481-488.

Benfenati, E., et al. "Predicting log*P* of pesticides using different software." *Chemosphere* 53.9 (2003): 1155-1164.

Benner, Steven A., Andrew D. Ellington, and Andreas Tauer. "Modern metabolism as a palimpsest of the RNA world." *Proceedings of the National Academy of Sciences* 86.18 (1989): 7054-7058.

Bergin, E.A., Aikawa, Y., Blake, G.A., and van Dishoeck, E.F. "The chemical composition of protoplanetary disks." *Protostars and Planets V*, edited by B. Reipurth, D. Jewitt, and K. Keil. University of Arizona Press, Tucson (2007).

Bernal, John Desmond. "The physical basis of life." (1951).

Bigelow, Charles C. "On the average hydrophobicity of proteins and the relation between it and protein structure." *Journal of theoretical biology* 16.2 (1967): 187.

Blum, Lorenz C., and Jean-Louis Reymond. "970 million druglike small molecules for virtual screening in the chemical universe database GDB-13." *J. Am. Chem. Soc* 131.25 (2009): 8732-8733.

Bohacek, Regine S., Colin McMartin, and Wayne C. Guida. "The art and practice of structure‐based drug design: A molecular modeling perspective."*ChemInform* 27.17 (1996).

Brooks, Dawn J., et al. "Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code."*Molecular Biology and Evolution* 19.10 (2002): 1645-1655.

Caporaso, J. Gregory, Michael Yarus, and Rob Knight. "Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code." *Journal of molecular evolution* 61.5 (2005): 597-607.

Caspi, Ron, et al. "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." *Nucleic acids research* 36.suppl 1 (2008): D623-D631.

Chambers, I., et al. "The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA." *The EMBO journal* 5.6 (1986): 1221.

Chyba, Christopher, and Carl Sagan. "Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life." (1992): 125-132.

Cleaves, H. James, et al. "A reassessment of prebiotic organic synthesis in neutral planetary atmospheres." *Origins of Life and Evolution of Biospheres* 38.2 (2008): 105-115.

Cleaves, H. James. "The origin of the biologically coded amino acids." *Journal of theoretical biology* 263.4 (2010): 490-498.

Cohn, E.J. and Edsall, J.T. "Proteins, Amino Acids, and Peptides", Reinhold, New York (1943).

Commans, Stephane, and August Böck. "Selenocysteine inserting tRNAs: an overview." *FEMS microbiology reviews* 23.3 (2006): 335-351.

Connerney, J. E. P., et al. "Magnetic lineations in the ancient crust of Mars." *Science* 284.5415 (1999): 794-798.

Crick, Francis H. C. "Codon-anticodon pairing: the wobble hypothesis." *J. mol. Biol* 19.2 (1966): 548-555.

Crick, Francis H. C. "An error in model building." [Editorial letter] *Nature* (1967): 798-798.

Crick, Francis H. C. "The origin of the genetic code." *Journal of molecular biology* 38.3 (1968): 367-379.

Cronin, J. R., G. W. Cooper, and S. Pizzarello. "Characteristics and formation of amino acids and hydroxy acids of the Murchison meteorite." *Advances in Space Research* 15.3 (1995): 91-97.

Cullmann, Georges, and Jean-Michel Labouygues. "Noise immunity of the genetic code." *Biosystems* 16.1 (1983): 9-29.

Cullmann, Georges, and Jean-Michel Labouygues. "The logic of the genetic code." *Mathematical Modelling* 8 (1987): 643-646.

Daviter, T., K. B. Gromadski, and M. V. Rodnina. "The ribosome's response to codon–anticodon mismatches." *Biochimie* 88.8 (2006): 1001-1011.

Denton, Michael J., Craig J. Marshall, and Michael Legge. "The protein folds as platonic forms: new support for the pre-Darwinian conception of evolution by natural law." *Journal of Theoretical Biology* 219.3 (2002): 325-342.

Des Marais, David J., et al. "The NASA astrobiology roadmap." *Astrobiology*8.4 (2008): 715-730.

Di Giulio, Massimo, and Mario Medugno. "Physicochemical optimization in the genetic code origin as the number of codified amino acids increases." *Journal of molecular evolution* 49.1 (1999): 1-10.

Di Giulio, Massimo. "The coevolution theory of the origin of the genetic code." *Journal of molecular evolution* 48.3 (1999): 253-254.

Di Giulio, Massimo. "The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA)." *Journal of theoretical biology* 240.3 (2006): 343-352.

Di Giulio, Massimo. "The origin of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analyses tautologous." *Journal of Theoretical Biology* 208.2 (2001): 141-144.

DiGiulio, Massimo. "Genetic Code Origin: Are the Pathways of Type Glu-tRNA$^{Gln}$ -> Gln-tRNA$^{Gln}$ Molecular Fossils or Not?." *Journal of molecular evolution* 55.5 (2002): 616-622.

Dobson, Christopher M. "Chemical space and biology." *Nature* 432.7019 (2004): 824-828.

Dragovich, Branko, and Alexandra Dragovich. "p-Adic Modelling of the Genome and the Genetic Code." *The Computer Journal* 53.4 (2010): 432-442.

Einstein, A, in *Festschrift fur Aurel Stodola*, E. Honegger, Ed. Orell Fussli Verlag, Zurich (1929): 126.

Ellington, Andrew D., and Jack W. Szostak. "In vitro selection of RNA molecules that bind specific ligands." *Nature* 346.6287 (1990): 818-822.

Elsila, Jamie E., Daniel P. Glavin, and Jason P. Dworkin. "Cometary glycine detected in samples returned by Stardust." *Meteoritics & Planetary Science* 44.9 (2009): 1323-1330.

Erives, Albert. "A Model of Proto-Anti-Codon RNA Enzymes Requiring l-Amino Acid Homochirality." *Journal of molecular evolution* 73.1 (2011): 10-22.

Feng, Liang, et al. "Aminoacyl-tRNA synthesis by pre-translational amino acid modification." *RNA biology* 1.1 (2004): 15-19.

Ferris, J. P., et al. "HCN: a plausible source of purines, pyrimidines and amino acids on the primitive Earth." *Journal of molecular evolution* 11.4 (1978): 293-311.

Figureau, A. "Information theory and the genetic code." *Origins of Life and Evolution of Biospheres* 17.3 (1987): 439-449.

Figureau, A. "Optimization and the genetic code." *Origins of Life and Evolution of Biospheres* 19.1 (1989): 57-67.

Figureau, A., and M. Pouzet. "Genetic code and optimal resistance to the effects of mutations." *Origins of Life and Evolution of Biospheres* 14.1 (1984): 579-588.

Fitch, W. M., and K. Upper. "The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code." *Cold Spring Harbor symposia on quantitative biology*. Vol. 52. Cold Spring Harbor Laboratory Press, (1987).

Flynn, George J., et al. "Elemental compositions of comet 81P/Wild 2 samples collected by Stardust." *Science* 314.5806 (2006): 1731-1735.

Freeland, Stephen J., and Laurence D. Hurst. "The genetic code is one in a million." *Journal of molecular evolution* 47.3 (1998): 238-248.

Freeland, Stephen J., Tao Wu, and Nick Keulmann. "The case for an error minimizing standard genetic code." *Origins of Life and Evolution of Biospheres*33.4 (2003): 457-477.

Freeland, Stephen J. "Terrestrial' Amino Acids and their evolution" in *Amino Acids, Peptides and Proteins within Organic Chemistry*, Vol. 1 (ed. A. B. Hughes), Wiley VCH (2009).

Fujishima, Kosuke, et al. "Sequence evidence in the archaeal genomes that tRNAs emerged through the combination of ancestral genes as 5′ and 3′ tRNA halves." *PLoS One* 3.2 (2008): e1622.

Galton, Francis. "Vox populi." *Nature* 75 (1907): 450-451.

Gilbert, Walter. "Origin of life: The RNA world." *Nature* 319.6055 (1986).

Gilson, Michael K., and Barry H. Honig. "Calculation of electrostatic potentials in an enzyme active site." *Nature* 330.6143 (1987): 84-86.

Golynskiy, Misha V., et al. "In Vitro Evolution of Enzymes." *Enzyme Engineering*. Humana Press, 2013. 73-92.

Gombar, V. K. "Reliable assessment of log P of compounds of pharmaceutical relevance." *SAR and QSAR in Environmental Research* 10.4 (1999): 371-380.

Goesmann, Fred, et al. "COSAC onboard Rosetta: a bioastronomy experiment for the short-period comet 67P/Churyumov-Gerasimenko." *Astrobiology* 5.5 (2005): 622-631.

Grantham, R. "Amino acid difference formula to help explain protein evolution." *Science (New York, NY)* 185.4154 (1974): 862.

Gugisch, Ralf, et al. "History and Progress of the Generation of Structural Formulae in Chemistry and its Applications." *Communications in Mathematical and in Computer Chemistry/MATCH* 58.2 (2007): 239-280.

Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C., Wassermann, A. "MOLGEN 5.0 Reference Guide" In http://molgen.de/documents/manual50.pdf (2009).

Haar, Lester. *NBS/NRC steam tables*. CRC (1984).

Haig, David, and Laurence D. Hurst. "A quantitative measure of error minimization in the genetic code." *Journal of molecular evolution* 33.5 (1991): 412-417.

Hatfield, Dolph L., and Vadim N. Gladyshev. "How selenium has altered our understanding of the genetic code." *Molecular and cellular biology* 22.11 (2002): 3565-3576.

Hernández-Montes, Georgina, et al. "The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution." *Genome Biol* 9.6 (2008): R95.

Higgs, Paul G., and Ralph E. Pudritz. "A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code." *Astrobiology* 9.5 (2009): 483-490.

Hinds, David A., and Michael Levitt. "From structure to sequence and back again." *Journal of molecular biology* 258.1 (1996): 201-210.

Hopfield, John J. "Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading." *Proceedings of the National Academy of Sciences* 75.9 (1978): 4334-4338.

Horowitz, N. H., G. L. Hobby, and Jerry S. Hubbard. "Viking on Mars: the carbon assimilation experiments." *Journal of Geophysical Research* 82.28 (1977): 4659-4662.

Ibba, Michael, Alan W. Curnow, and Dieter Söll. "Aminoacyl-tRNA synthesis: divergent routes to a common goal." *Trends in biochemical sciences* 22.2 (1997): 39-42.

Ilardo, Melissa A. "Thawing the Frozen Accident: Variation in the Genetic Code." *Junior Independent Work* (2010).

Irvine, Doug, Craig Tuerk, and Larry Gold. "SELEXION: Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis." *Journal of molecular biology* 222.3 (1991): 739-761.

Johnson, Adam P., et al. "The Miller volcanic spark discharge experiment." Science 322.5900 (2008): 404-404.

Johnson, David BF, et al. "RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites." *Nature chemical biology* 7.11 (2011): 779-786.

Jukes, T.H. "Arginine as an evolutionary intruder into protein synthesis." *Biochem. Biophys. Res. Commun.* 53 (1973): 709-714.

Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* 28.1 (2000): 27-30.

Kasting, James F. "Earth's early atmosphere." *Science* (1993): 920-920.

Kauzmann, W. "Of Protein Denaturation." *Advances in protein chemistry* 14 (1959): 1.

Kawashima, Shuichi, Hiroyuki Ogata, and Minoru Kanehisa. "AAindex: amino acid index database." *Nucleic Acids Research* 27.1 (1999): 368-369.

Kerridge, John F., et al. "An exobiological strategy for Mars exploration." *NASA Spec. Publ., NASA SP* 530 (1995): 55.

Klein, Harold P. "The Viking biological experiments on Mars." *Icarus* 34.3 (1978): 666-674.

Knight, Robin D., Stephen J. Freeland, and Laura F. Landweber. "Rewiring the keyboard: evolvability of the genetic code." *Nature Reviews Genetics* 2.1 (2001): 49-58.

Kolodny, Rachel, et al. "On the Universe of Protein Folds." *Annual Review of Biophysics* 42.1 (2013).

Koonin, Eugene V., and Artem S. Novozhilov. "Origin and evolution of the genetic code: the universal enigma." *IUBMB life* 61.2 (2009): 99-111.

Krzycki, Joseph A. "The direct genetic encoding of pyrrolysine." *Current opinion in microbiology* 8.6 (2005): 706-712.

Kvenvolden, Keith, et al. "Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite." (1970): 923-926.

Ladunga, Istvan, and Randall F. Smith. "Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties." *Protein engineering* 10.3 (1997): 187-196.

Lazcano, Antonio, and Stanley L. Miller. "On the origin of metabolic pathways."*Journal of molecular evolution* 49.4 (1999): 424-431.

Leaver-Fay, Andrew, et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." *Methods Enzymol* 487 (2011): 545-574.

Leinfelder, Walfred, et al. "Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine." (1988): 723-725.

Levitt, Michael. "Nature of the protein universe." *Proceedings of the National Academy of Sciences* 106.27 (2009): 11079-11084.

Liu, Chang C., and Peter G. Schultz. "Adding new chemistries to the genetic code." *Annual review of biochemistry* 79 (2010): 413-444.

Liu, Chang C., and Peter G. Schultz. "Adding new chemistries to the genetic code." *Annual review of biochemistry* 79 (2010): 413-444.

Lloyd, D. G., Golfis, G., Knox, A. J., Fayne, D., Meegan, M. J., & Oprea, T. I. "Oncology exploration: charting cancer medicinal chemistry space." *Drug discovery today* 11.3 (2006): 149-159.

Longstaff, David G., et al. "A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine." *Proceedings of the National Academy of Sciences* 104.3 (2007): 1021-1026.

Mathew, Damien C., and Zaida Luthey-Schulten. "On the physical basis of the amino acid polar requirement." *Journal of molecular evolution* 66.5 (2008): 519-528.

McKay, Christopher P., and Wanda L. Davis. "Duration of liquid water habitats on early Mars." *Icarus* 90.2 (1991): 214-221.

Meringer, Markus. "Structure Enumeration and Sampling." *Handbook of Chemoinformatics Algorithms* 33 (2010): 233.

Meringer M., Cleaves H.J. and Freeland S.J., “Beyond Terrestrial Biology: Charting the Chemical Universe of a-Amino Acid Structures,” *Jounral of Chemical Informatics*, in review.

Miller, Stanley L. "A production of amino acids under possible primitive earth conditions." *Science* 117.3046 (1953): 528-529.

Miller, Stanley L., and C. L. Harold. "Organic Compound Synthes on the Primitive Eart." *Science* 130 (1959): 251.

Moult, John, et al. "Critical assessment of methods of protein structure prediction (CASP)—round IX." *Proteins: Structure, Function, and Bioinformatics* 79.S10 (2011): 1-5.

Moura, Gabriela R., et al. "Species-specific codon context rules unveil non-neutrality effects of synonymous mutations." *PloS one* 6.10 (2011): e26817.

Nakai, Kenta, Akinori Kidera, and Minoru Kanehisa. "Cluster analysis of amino acid indices for prediction of protein structure and function." *Protein engineering* 2.2 (1988): 93-100.

Navarro-González, Rafael, et al. "The limitations on organic detection in Mars-like soils by thermal volatilization–gas chromatography–MS and their implications for the Viking results." *Proceedings of the National Academy of Sciences* 103.44 (2006): 16089-16094.

Noren, Christopher J., et al. "A general method for site-specific incorporation of unnatural amino acids into proteins." *Science* 244.4901 (1989): 182-188.

Oparin, Aleksandr Ivanovich. *Origin of life*. Dover Publications (1938).

Osawa, Syozo, and Thomas H. Jukes. "Codon reassignment (codon capture) in evolution." *Journal of molecular evolution* 28.4 (1989): 271-278.

Parker, Eric T., et al. "Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment." *Proceedings of the National Academy of Sciences* 108.14 (2011): 5526-5531.

Patel, Shona C., and Michael H. Hecht. "Directed evolution of the peroxidase activity of a de novo-designed protein." *Protein Engineering Design and Selection* 25.9 (2012): 445-452.

Pelc, S. R., and M. G. Welton. "Stereochemical relationship between coding triplets and amino-acids." *Nature* 209.5026 (1966): 868.

Pizzarello, Sandra, and Everett Shock. "The organic composition of carbonaceous meteorites: The evolutionary story ahead of biochemistry." *Cold Spring Harbor perspectives in biology* 2.3 (2010).

Philip, Gayle K., and Stephen J. Freeland. "Did Evolution Select a Nonrandom “Alphabet” of Amino Acids?." *Astrobiology* 11.3 (2011): 235-240.

Plankensteiner, Kristof, Hannes Reiner, and Bernd M. Rode. "Amino acids on the rampant primordial Earth: electric discharges and the hot salty ocean." *Molecular diversity* 10.1 (2006): 3-7.

Prialnik, Dina, et al. "Thermal and chemical evolution of comet nuclei and Kuiper belt objects." *Space Science Reviews* 138.1-4 (2008): 147-164.

Reymond, Jean-Louis, and Mahendra Awale. "Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database." *ACS Chemical Neuroscience* (2012).

Sandford, Scott A., et al. "Organics captured from comet 81P/Wild 2 by the Stardust spacecraft." *Science* 314.5806 (2006): 1720-1724.

Schlesinger, Gordon, and Stanley L. Miller. "Prebiotic synthesis in atmospheres containing CH4, CO, and CO2." *Journal of molecular evolution*19.5 (1983): 383-390.

Schomburg, Ida, Antje Chang, and Dietmar Schomburg. "BRENDA, enzyme data and metabolic information." *Nucleic acids research* 30.1 (2002): 47-49.

Sella, Guy, and David H. Ardell. "The coevolution of genes and genetic codes: Crick's frozen accident revisited." *Journal of molecular evolution* 63.3 (2006): 297-313.

Sephton, Mark A. "Organic compounds in carbonaceous meteorites." *Natural Product Reports* 19.3 (2002): 292-311.

Snyder, L. E., et al. "A rigorous attempt to verify interstellar glycine." *The Astrophysical Journal* 619.2 (2008): 914.

Sonneborn, T. M. "Degeneracy of the genetic code: extent, nature, and genetic implications." *Evolving genes and proteins*. Academic Press, New York (1965): 377-397.

Srinivasan, Gayathri, Carey M. James, and Joseph A. Krzycki. "Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA." *Science* 296.5572 (2002): 1459-1462.

Srinivasan, Gayathri, Carey M. James, and Joseph A. Krzycki. "Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA."*Science* 296.5572 (2002): 1459-1462.

Steel, Robert GD, and James Hiram Torrie. *Principles and procedures of statistics, a biometrical approach*. No. Ed. 2. McGraw-Hill Kogakusha, Ltd., 1980.

Taylor, F. J. R., and D. Coates. "The code within the codons." *Biosystems*22.3 (1989): 177-187.

Taylor, G. Jeffrey, et al. "Bulk composition and early differentiation of Mars."*Journal of Geophysical Research: Planets (1991–2012)* 111.E3 (2006).

ten Kate, Inge L. "Organics on Mars?." *Astrobiology* 10.6 (2010): 589-603.

Tian, Feng, et al. "A hydrogen-rich early Earth atmosphere." *Science* 308.5724 (2005): 1014-1017.

Tumbula, Debra L., et al. "Domain-specific recruitment of amide amino acids for protein synthesis." *Nature* 407.6800 (2000): 106-110.

Urey, Harold C., and Serge A. Korff. "The planets: their origin and development." *Physics Today* 5 (1952): 12.

Uy, Rosa, and Finn Wold. "Posttranslational covalent modification of proteins." *Science* 198.4320 (1977): 890-896.

van der Waterbeemd, H., H. Karajiannis, and N. El Tayar. "Lipophilicity of amino acids." *Amino Acids* 7.2 (1994): 129-145.

Wächtershäuser, Günter. "Towards a reconstruction of ancestral genomes by gene cluster alignment." *Systematic and applied microbiology* 21.4 (1998): 473-477.

Walker, James CG, and Peter Brimblecombe. "Iron and sulfur in the pre-biologic ocean." Precambrian Research 28.3 (1985): 205-222.

Weber, Arthur L., and Stanley L. Miller. "Reasons for the occurrence of the twenty coded protein amino acids." *Journal of Molecular Evolution* 17.5 (1981): 273-284.

Weberndorfer, Günter, Ivo L. Hofacker, and Peter F. Stadler. "On the evolution of primitive genetic codes." *Origins of Life and Evolution of the Biosphere* 33.4-5 (2003): 491-514.

White, Harold B. "Coenzymes as fossils of an earlier metabolic state." *Journal of Molecular Evolution* 7.2 (1976): 101-104.

Woese, C. Rt. "On the evolution of the genetic code." Proceedings of the National Academy of Sciences of the United States of America 54.6 (1965): 1546.

Woese, C. R., et al. "The molecular basis for the genetic code." *Proceedings of the National Academy of Sciences of the United States of America* 55.4 (1966): 966.

Woese CR. The genetic code: the molecular basis for genetic expression. Harper & Row, New York (1967).

Woese, Carl R. "Evolution of the genetic code." *Naturwissenschaften* 60.10 (1973): 447-459.

Wolman, Yecheskel, William J. Haverland, and Stanley L. Miller. "Nonprotein amino acids from spark discharges and their comparison with the Murchison meteorite amino acids." *Proceedings of the National Academy of Sciences*69.4 (1972): 809-811.

Wong, J. "Coevolution theory of the genetic code at age thirty." *BioEssays*27.4 (2005): 416-425.

Wong, J. Tze-Fei. "A co-evolution theory of the genetic code." *Proceedings of the National Academy of Sciences of the United States of America* 72.5 (1975): 1909.

Wong, Jeffrey Tze-Fei. "Question 6: coevolution theory of the genetic code: a proven theory." *Origins of Life and Evolution of Biospheres* 37.4 (2007): 403-408.

Wrabl, James O., and Nick V. Grishin. "Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization." *Proteins: Structure, Function, and Bioinformatics* 61.3 (2005): 523-534.

Xu, Xue-Ming, et al. "Biosynthesis of selenocysteine on its tRNA in eukaryotes." *PLoS biology* 5.1 (2006): e4.

Yarus, Michael. "A specific amino acid binding site composed of RNA." *Science* 240.4860 (1988): 1751.

Yarus, M. "An RNA-amino acid complex and the origin of the genetic code." *New Biol* 3.2 (1991): 183-189.

Yarus, Michael. "Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin." *Journal of molecular evolution* 47.1 (1998): 109-117.

Yarus, Michael, Jeremy Joseph Widmann, and Rob Knight. "RNA–amino acid binding: A stereochemical era for the genetic code." *Journal of molecular evolution* 69.5 (2009): 406-429.

Yuan, Jing, et al. "RNA-dependent conversion of phosphoserine forms selenocysteine in eukaryotes and archaea." *Proceedings of the National Academy of Sciences* 103.50 (2006): 18923-18927.

Zeng, Xiang, et al. "Pyrococcus CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life." *The ISME journal* 3.7 (2009): 873-876.

Zhang, Yan, et al. "Pyrrolysine and selenocysteine use dissimilar decoding strategies." *Journal of Biological Chemistry* 280.21 (2005): 20740-20751.

Zuckerkandl, Emile, and Linus Pauling. "Evolutionary divergence and convergence in proteins." *Evolving genes and proteins* 97 (1965): 166.