A Computational Investigation on Properties of the Standard Genetic Code with Implications for its Origin

A Computational Investigation on Properties of the Standard Genetic Code with Implications for its Origin

By:

Gregory Di Sanza

Supervisor:

Dr. Jonathon Stone

A Thesis Submitted to McMaster University SGS
for Partial Completion of the Requirements for a
Master of Science Degree

McMaster University
September 2016
© Copyright by Gregory Di Sanza 2016

Master of Science (2016) McMaster University
Department of Biology Origins Institute: Collaborative
Astrobiology Graduate Program

Title: A Computational Investigation and Review of the Structural Origins of the Genetic Code

Author: Gregory Di Sanza, H.BSc.

Supervisor: Dr. Jonathon Stone

Number of Pages: 133

Abstract

The amino acid blueprints for many essential biomolecules, including proteins like immunoglobulins, insulin, metabolic enzymes, electron transporters, and structural molecules, are encoded within the standard genetic code (SGC). These blueprints are stored in nucleic acids such as DNA or RNA and translated into amino acids by tRNA and ribosomes. Investigations into the origins of and principles guiding SGC structure have been performed since the 1960s, with a majority revealing that amino acids with similar physicochemical properties are to assigned triplet codons with similar nucleotides. A significant amount of degeneracy is seen particularly at the third position of most codons, leading to hypotheses that buffering against transcriptional and translational error was the driving force behind SGC structure. In this study, we review some prominent theories on the origins of the triplet codon system and possible mechanisms for how those codons became assigned to amino acids and examine briefly non-bonding forces and their effects on amino acid interactions.

A computational program was developed to run simulations comparing the SGC to other, hypothetical genetic codes (HGCs), generated by shuffling amino acid identities among codon blocks and performing error measure calculations. The SGC was ranked against the HGCs for 54 properties. Shuffling of stop codon positions was implemented for only the second time with such analyses. When stop codons were included in error measure calculations, minimum, zero, and maximum penalties were assigned to nonsense mutations. Long-range non-bonded energy was found to be the most conserved property when accounting for stop codon effects.

Acknowledgements

Firstly, I wish to acknowledge the teachings, guidance, encouragement, and wisdom I have received under the tutelage of Dr. Jonathon Stone. Among the student and faculty bodies of McMaster University, may he be remembered for his patience, modesty, cordial manner, and generosity. Without Dr. Stone's continuous support, beginning in my undergraduate career and persisting until present day, the opportunity to achieve a graduate degree would not have been possible. If any professor deserves substantial funding grants, and high impact publications, and recognition for always having the students best interests at heart, it is him. Gratitude is also extended to Tarushika Vasanthan (PhD. Candidate), whose constant support, advice, and guidance in things both science and mundane helped make my graduate studies so memorable. Working alongside her has been both an honour and a privilege.

Secondly, I would also like to thank my parents, Tony Di Sanza and Virginia Allega-Di Sanza, and brother, Kevin Di Sanza. Their financial & material support, encouragement, love and belief in my abilities has not only brought forth my greatest potential, but also given me the fortitude, will, and means to reach my goals. My debt to them is so much greater than they will ever know.

Thirdly, I thank the Origins Institute and collaborative Astrobiology Graduate Program for providing the funding to complete my project.

Finally, I extend thanks to all the friends, both young and old, that I have made at McMaster University over the years. Many became like brothers and sisters to me, deepening my connection to this campus and providing treasured memories and experiences.

Table of Contents

List of Figures

Degagne 2015:Figure 10 - Ranking for SGC relative to three sets of 100000 HGCs that correspond to different theories on the origin of the genetic code. Polar requirement was the property analyzed and mean square error used as the distance metric.

Goodarzi et al. 2004:Figure 2 - Distribution of ranks for 2×10^9 HGCs with the SGC rank indicated by an arrow.

Goodarzi et al. 2004:Figure 3 - Z-values plotted against varying K-values for $PAM_{74\text{-}100}$ matrix and mutation matrix.

Goodarzi et al. 2004:Figure 5 - Distribution scores of 1.2×10^9 HGCs when accounting for nonsense mutations in both $PAM_{74\text{-}100}$ matrix and mutation matrix.

Street and Mayo 1999:Figure 1 - Propensity for formation of β-pleated sheets among amino acids.

Grimme 2004:Figure 4 - Optimized base pair structures and conformations among aromatic molecules and nucleotide base pairs.

List of Tables

List of Abbreviations

DNA: deoxyribonucleic acid

RNA: ribonucleic acid

mRNA: messenger ribonucleic acid

tRNA: transfer ribonucleic acid

HGC: hypothetical genetic code

PGC: primordial genetic code

SGC: standard genetic code

aaRS: aminoacyl tRNA synthetase

N: any nucleotide, as in A, G, C, or T/U (e.g., codon GCN, where N may be any nucleotide)

Introduction

One of the many topics on which astrobiologists focus research is the origin and evolution of a primordial genetic code (PGC) into the standard genetic code (SGC) present in almost all organisms. Several events likely occurred in producing the SGC, including the synthesis of nucleotides from ribose sugars, carbon and phosphate groups; the aggregation of nucleotides into strands of RNA; the emergence of codons in a triplet code; and the origin of tRNA with anticodons and charging sites, and tRNA synthetases that can bind amino acids to tRNA. Many studies have analyzed error rates at codon positions, codon block arrangements, amino acid frequencies, why similar amino acids are assigned to similar codons, and optimization in the SGC organization. The majority of these investigations involved computer simulations, in which the SGC was compared to hypothetical genetic codes (HGCs), although only a small minority of studies accounted for effects imparted by shuffling termination, or 'stop', codon positions.

Whereas several theories on the origin of the SGC have been proposed, three of particular interests are the coevolution theory (Wong 1975, 2005), first-principles theory (Trifonov 2000, 2004), and four-column theory (Higgs 2009). The coevolutionary theory suggests that amino acids became assigned to codons that previously had encoded precursor amino acids in biosynthetic pathways that resulted in the synthesis of the newly assigned amino acids (i.e., amino acids became assigned to codons that belonged to their biosynthetic precursors). The four-column theory postulates that amino acids were added to a PGC comprising redundant codon sets, or columns, according to similarities between their physico-chemical properties and positive selection for increased complexity and

diversity of proteins that can be produced from a larger amino acid repertoire. This review provides an examination of these theories.

CHAPTER 1

1.1 Coevolution Theory

Wong (1975) postulated that the mechanisms of prebiotic synthesis were insufficient to form all 20 standard amino acids, suggesting that a significant portion of them must have had their origins in the coevolving metabolic pathways of amino acid biosynthesis. As early amino acids would have been utilized in newly evolving metabolic pathways, other amino acids could have been synthesized from the available substrates. The basis of the prevailing structure in the SGC is the result of precursor-product relationships between amino acids during biosynthesis. The codon domains of these precursor-product pairs probably differentiated by single nucleotide changes, allowing for pairs to remain contiguous. New amino acids were assigned to codons that belonged to their biosynthetic precursors.

Wong (1975) stated that, if amino acid pairs were associated strongly in precursor-product relationships (e.g., precursors could be converted to products in relatively few enzymatic steps) in biosynthetic pathways but failed to be assigned to contiguous codon domains, then the coevolution theory would be unsustainable. To show that this is not the case, relationships between precursor-product pair conversions among the amino acids known at the time were examined (Wong 1975:Figure on page 1909). The one that required particular explanation was the conversion of Asp to Lys. Wong (1975) explained that the conversion could take place through two different pathways.

One pathway is used by prokaryotic organisms and is called the diaminopimelate pathway, which converts aspartic acid into aspartyl phosphate and then into an aspartic semialdehyde that branches off to synthesize lysine and diaminopimelic acid via aspartic semialdehyde dehydrogenase or aspartic kinase enzymes (Gilvarg 1962). The other pathway is used by eukaryotes (specifically certain fungi) and is called the α-aminoadipate pathway, which converts homocitrate into homoisocitrate through a dehydrogenase and then transfers a nitrogen atom from glutamate onto homoisocitrate, forming an α-aminoadipate, which eventually is transformed into lysine by enzymatic activity (Strassman and Weinhouse 1952).

Wong (1975) also argued that most of the codons for precursor amino acids are indeed contiguous with codons for product amino acids, the exceptions being the pairs Glu-Pro, Glu-Arg, Asp-Thr, and Asp-Lys. During the early stages of codon assignment in a PGC, the CAA & CAG codons and the AAU & AAC codons likely were assigned to the Glu codon block and Asp codon block, respectively, as long as prebiotic synthesis of Gln from Glu and Asn from Asp did not occur. This proposal concerning the two dicarboxylic amino acids (Glu and Asp) is sufficient to eliminate all of the non-contiguities between the precursor amino acids and their biosynthetic products. Adopting this assumption, Wong (1975:Figure 1) constructed a map of the genetic code wherein codon domains of precursor-product and interchangeable amino acid pairs were distinct from each other only by single altered nucleotides. These relationships were taken as evidence that amino acid biosynthesis and codon assignments are correlated strongly.

The coevolution theory also proposes that pathways used in prebiotic synthesis eventually gave rise to pathways used in enzymatic biosynthesis. These pathways are

interpreted as having co-evolved, where prebiotically formed amino acids were used later in biosynthetic pathways to generate later amino acids, with the distribution of triplet codons to product amino acids being determined by precursor amino acids. Randomness tests were conducted to show that correlations between codon assignments and amino acid precursor-product pairs were not due to chance events. This was achieved via two methods: analysis of the extensive contiguities between codon assignments in product amino acids vs. precursor amino acids and calculating the probability that three out of any seven amino acid pairs actually are sibling pairs (e.g., product amino acids that were derived from the same parent).

Analyzing contiguities between codon assignments and amino acid precursor-product pairs involves two groups of codon triplets. For any precursor codon triplets, one group will be contiguous with the precursors (a) while the other group will be non-contiguous with the precursors (b). If the product amino acid that was synthesized from the precursor has n codons assigned to it, then a probability P that x among those codons are contiguous with a different precursor amino acid can be calculated with a hypergeometric distribution. $P = \sum_{x}^{n} \frac{a!}{(a-x)!x!} \cdot \frac{b!}{(b-n+x)!(n-x)!} \cdot \frac{(a+b-n)!n!}{(a+b)!}$.

Following the method that was used by Fisher (1950), probabilities for precursor-product codon contiguities were calculated for 8 amino acid pairs. The resulting chi-square value corresponded to a probability of 0.0002 that the 8 precursor-product amino acid contiguities arose by chance. However, the Phe-Tyr and Val-Leu pairs were expected to have been produced from a shared biosynthetic pathway rather than being true precursor-product pairs, so the calculation was repeated with these pairs excluded. That chi-square

value still yielded a 0.0075 probability that observed amino acid contiguities arose by chance. Other possible contiguities that were excluded because they would have decreased the aggregate probability were: Glu-Pro, Asp-Thr, Asp-Lys, Thr-Met, Ala-Ser-Gly, and Glu-Asp-Ala.

Calculating the probability that three out of any seven amino acid pairs are sibling pairs rather than a true precursor-product pair involved considering the biosynthetic sibling amino acids for the Glu, Asp, Ser, and Thr families: Gln, Pro, and Arg are siblings to the Glu amino acid family; Asn, Thr, and Lys are siblings to the Asp amino acid family; Cys and Trp are siblings to the Ser family; and Ile and Met are siblings of the Thr family. In the SGC, 7 pairs of amino acids share the same nucleotides at the 1st and 2nd positions in their codons: Asn-Lys, Cys-Trp and Ile-Met are the three sibling pairs; the Gln-His pair is a precursor-product and the Asp-Glu pair can be either siblings or a precursor-product, leaving Phe-Leu & Ser-Arg as completely unrelated amino acid pairs. The 20 amino acids of the genetic code can be organized into 190 possible amino acid pairs, with the 4 families of sibling amino acids yielding a total of 8 pairs of sibling amino acids. Applying the hypergeometric distribution equation,

$$P = \sum_{x}^{n} \frac{a!}{(a-x)!x!} \bullet \frac{b!}{(b-n+x)!(n-x)!} \bullet \frac{(a+b-n)!n!}{(a+b)!}$$

, results in a 0.00161 probability that up to 3 of any 7 pairs are true sibling pairs due to chance. Even when the Ile-Met pair is not counted as a sibling, the probability of the relationships observed being due to chance only increases to 0.0224. However, Wong also noted that Asp and Glu can be considered as a sibling pair, decreasing the probability back to 0.00161. Overall, the probability of amino acid pairs that share the same nucleotides at the 1st and 2nd codon

position indicate that amino acid biosynthesis and codon distribution are correlated strongly (Wong 1975:Table 2).

Based on prebiotic synthesis experiments (Miller 1953, Palm and Calvin 1962, Harada et al. 1964), Wong (1975) placed Glu, Asp, Gly, Ala, and Ser as major members in the first amino acids, with Phe-Tyr and Val-Leu as minor members. These arrangements of major and minor members were based on which amino acids most likely would have existed in greatest abundance during prebiotic conditions. Wong (1975) argued that the interconversions and intermediates within carbohydrate and lipid metabolism, the addition of Gly, Asp, and Gln into developing nucleobase synthesis pathways, and the presence of molecules with α-transaminase activity (that utilize Glu and Ala) would have been sufficient for the biosynthesis of almost all later amino acids. However, due to the inherent lack of complex proteins or molecules in the primordial environment, the earliest prebiotic reaction pathways likely proceeded uncatalyzed. As faster reaction rates became favoured, these early pathways began to incorporate molecules with enzymatic activity into their reaction pathways. For example, the biosynthetic pathway for Pro synthesis includes the spontaneous cycling of glutamyl-γ-semialdehyde into Δ1-pyrroline-5-carboxylate, which may be a relic of prebiotic synthesis that incorporated enzymatic activity into a modern-day pathway (Vogel and Davis 1952).

Wong (1975) also proposed an answer to the question ‘why were precursor-product amino acids assigned to contiguous codons in the first place?’ Product amino acids were assigned to codons that were previously assigned to precursor amino acids. A more detailed solution was deduced from a map (Wong 1975:Figure 1). In the Glu & Gln

and Asp & Asn codon blocks, CAA and CAG initially were codons for Glu and later were passed to Gln; AAU and AAC initially were codons for Asp and later were passed to Asn. The PGC therefore contained only a small set of amino acids that each individually occupied a contiguous codon domain. As precursor amino acids were subjected to emerging biosynthetic pathways that involved non-enzymatic catalysts, product amino acids were formed. These new product amino acids took codons from their biosynthetic precursor amino acids, with the precursor amino acids taking up neighbouring codons that had not yet been assigned to any specific amino acid. Three specific conditions would have allowed codon reassignments to have occurred, all involving a primordial type of tRNA called an adaptor molecule: (1) the structure of the product must have resembled the structure of the precursor strongly enough to have contended for binding sites on adaptor molecules; (2) while a precursor was bound to an adaptor molecule, it would have to have undergone a conversion reaction to form the product amino acid, taking on its precursor codons; (3) the prebiotic reaction pathway from precursor to product would have to have contained an intermediate molecule, which bound to an adaptor molecule for a precursor amino acid and then transformed to a product amino acid. Wong (1975) proposed that these conditions maintained codon domain contiguities with precursor-product pairs.

Four factors must be appreciated in searching for and identifying structural similarities between prokaryotic tRNA molecules and amino acids of the same biosynthetic family: (1) the adaptors that are yielded by precursors to products may be different than adaptors that precursor molecules retain; (2) complete structural variation is not required for adaptor molecules of different precursor amino acids; (3) structures of

tRNA molecules differ between prokaryotic and eukaryotic life forms, possibly with widespread divergent and convergent evolution occurring between primordial life forms and the emergence of prokaryotic organisms; (4) a 40-70% difference between nucleotide sequences of tRNA that can interact with multiple different amino acids can exist, which makes the reconstruction of an evolutionary map for tRNA challenging. The fourth factor suggests that tRNA comparisons based on general structure are insufficient for evolutionary analyses, so additional comparisons must be performed between specific regions, such as anticodon loops, where interactions between tRNA and codons occur.

Wong (1975:Table 3) reviewed examples of resemblances between anticodon stem-loops for different amino acid pairs that share biosynthetic pathways in *E. coli*, which are equivocal. Previous research (Folk et al. 1972) had shown that $tRNA_1^{Ser}$ and $tRNA^{Trp}$ contain common anticodon stem sequences that are present in $tRNA^{Gln}$, indicating that the stem loop similarities are not completely specific. Additionally (Yamada and Ishikura 1973; Ish-Horowicz and Clark 1973), even though $tRNA_1^{Ser}$ and $tRNA^{Trp}$ possess identical stem sequences, they differ from the stem sequence in $tRNA_3^{Ser}$ (which is C-U-C-C-C…G-G-G-A-G), indicating that the stem loop sequences are not completely non-specific; this is evidenced by the fact that $tRNA_1^{Ser}$ serves codons that are contiguous to the UGG codon of Trp but $tRNA_3^{Ser}$ does not. Although these similarities are neither totally specific nor totally non-specific, they are consistent with precursor amino acids yielding their codons to product amino acids. Wong (1975) proposed that tRNA anticodons evolved quite early in the coevolutionary era and operated as adaptor molecules for amino acids within protein synthesis.

The effects that other factors may have had on the assignment of codons to specific locations within the SGC also have been considered. These factors include codon plurality among amino acids and codon contiguity between amino acids that are chemically similar. Detrimental effects from errors in coding or excess mutation could have been mitigated by codon plurality and contiguity (Sonneborn 1965, Woese 1965). For example, when Asp was first used to synthesize Lys, many competing versions of the genetic code might have existed, with Lys placed in different domains that were contiguous with codons of Asp. The genetic code that assigned the codons AAA and AAG to Lys would have had the greatest fitness among any competing genetic codes, selecting for the genetic code where the Arg and Lys domains were adjacent to each other. Wong (1975) noted that, after the initial genetic code had been established, additional changes would have been lacking, likely due to the fact the precursor-product contiguities and sibling contiguities would have been maintained. This same conclusion was reached in separate studies that examined radical impacts that were caused by large changes to the SGC (Crick 1963, Hinegardner and Engelberg 1963).

Jukes (1971) proposed that amino acids that had been incorporated into a PGC earlier would have had more opportunities to have developed codon plurality than would have amino acids that were incorporated later. Late arrivals have fewer codons assigned to them, such as Met and Trp, with only one each. Amino acids that were synthesized after the SGC had been established (such as hydroxylysine or hydroxyproline) but not integrated into the genetic code itself could have been incorporated only into proteins through post-translational modification pathways. In spite of this, Wong (1975) noted that codon plurality on its own is insufficient to categorize amino acids, like Glu and Asp,

as early entrants in a PGC: the early entry of an amino acid to the PGC is not the exclusive factor that determines codon plurality. Wong (1975) claimed that the ability of amino acids to retain codon plurality depended on stasis, with codon and amino acid assignments being neutral selectively. Generally, however, earliness of amino acid incorporation determined codon plurality, whereas "evolutionary inertness" determined codon retention. As examples of codon retention, Leu and Arg seem to have not evolved or been altered by biosynthetic pathways and each retain up to 6 codons. Contrarily, Glu, despite being one of the earliest amino acids to have been added, has been subjected to evolution enough times to produce new families. This gives Glu low codon retention, with only two triplets assigned to it, but a considerable total of 16 codons are assigned to amino acids in the Glu family. The situation for Asp acid is similar, with 14 codons in the Asp family, but Asp, itself, retaining only two codons. Early incorporation and evolutionary inertness allowed some amino acids to obtain and preserve high codon plurality.

1.1.2 Coevolution Theory From a Modern Perspective: 4 Premises

Wong (2005) revisited the coevolution theory to reassess its main postulates after 30 years of new knowledge. The 4 central premises, amino acid biosynthesis requirement, pretran synthesis, biosynthetic imprints left on codon assignments, and mutability of amino acids within the code, were supported by new information and data. Wong (2005) also drew upon previous studies to decipher how much three different factors had shaped the SGC, based on their ability to reduce the number of possible genetic codes, N. Error minimization alone could account for the SGC as a 1-in-a-million genetic code (10^{-6}; Freeland and Hurst 1998). Stereochemical interaction based on the

binding of aptamers to amino acids results in 4 x 10^{-4} N-reduction, and amino acid biosynthesis resulted in 10^{-11} N-reduction (Freeland and Hurst 1998, Wong 1980, Knight et al. 2003). Overall, the relative input of amino acid biosynthesis: error minimization: stereochemical interaction occurred at a first order approximation ratio of 40000000:400:1. This initial estimate suggests that amino acid biosynthesis was by far the most dominant factor in guiding the evolution of the genetic code.

Wong (2005) briefly reviewed 4 general premises that have been used to explain the fact that the SGC occurs almost universally among organisms: amino acid biosynthesis, error minimization, stereochemical interactions, and codon expansion. Amino acid biosynthesis models propose that the various biosynthetic pathways that formed after the initial amino acids had been encoded are what directed the development of the SGC structure (Jukes 1973, Dillon 1973, Wong 1975, 1976, 1981, 1988, McClendon 1986, Danchin 1989, Di Giulio 1989, 1996, 2004, Di Giulio and Medugno 1998, Wachtershauser 1988, Taylor and Coates 1989, Edwards 1996, Budisa et al. 1999, Bermudez et al. 1999, Stevenson 2002, Cavalcanti et al. 2002, Klipcan and Safro 2004). Error minimization models state that the SGC was structured such that transcription and translation errors were minimized by configuring codons for structurally similar amino acids into adjacent positions in codon identity space (Sonneborn 1965, Woese et al. 1966, Jung 1978, Figureau 1989, Luo 1989, Lacey et al. 1992, Goldman 1993, Freeland and Hurst 1998, Freeland et al. 2003, Judson and Haydon 1999, Chiusano et al. 2000). Stereochemical interaction models refer to the stereochemistry that could have occurred between amino acids and their codons or tRNA anticodons, influencing positions in

codon identity space (Pelc 1965, Dunnill 1966, Melcher 1974, Hendry et al. 1981, Shimizu 1982, Yarus 1988, Knight et al. 2003, Seligmann and Amzallag 2002).

Finally, codon expansion models argue that triplet codons in a PGC were few in number, most triplets being unavailable or unused by early amino acids, with the total eventually growing in number to 64 as the number of encoded amino acids increased (Eigen and Schuster 1978, Jurka and Smith 1987, Delarue 1995, Davis 1999, Yang 2004, Guimaraes and Moreira 2004). In the 30-year revision of the coevolution theory, Wong (2005) noted that the assignment of codons to amino acids might have been determined by alterations to encoded amino acids. However, these chemical alterations were based on magnitudes of similarities in amino acid structures, in contrast to how numerous metabolic transformations actually take place. For example, the biosynthetic precursor to Arg is ornithine. These two amino acids could have contended with each other for attachment onto an ornithine-tRNA, leaving a biosynthetic imprint on the genetic code. The four premises are reviewed subsequently.

1.1.2.1 Premise 1: Amino Acid Biosynthesis

The first premise, amino acid biosynthesis, divides amino acids into three different categories: phase-1, phase-2, and phase-3. Phase-1 amino acids are those that appeared earliest due to the fact that they could have been synthesized by prebiotic synthesis alone. The atmosphere of the early Earth was probably only mildly reducing, so synthesis of amino acids from electrical discharge was unlikely. Delivery of high-energy protons to mixtures of carbon monoxide, nitrogen, and water, would have resulted in the prebiotic synthesis of phase-1 amino acids (Kobayashi et al. 1990, Kobayashi et al.

1998). Amino acids predicted to have been synthesized prebiotically, phase-1 constituents, were compared to amino acids that are observed to occur through irradiation of a carbon monoxide-nitrogen-water solution (Wong 2005:Table 1). None of the phase-2 amino acids could have been or can be synthesized through these means, and they accordingly are hypothesized to have appeared only after biosynthetic pathways had emerged to convert phase-1 amino acids into phase-2 amino acids. A study of 40 different factors that relate to the order of appearances of amino acids (Trifonov 2000) found a division between earlier appearing and later appearing amino acids that is consistent with the phase-1 and phase-2 amino acid separation on coevolution theory. The only exception to this was the phase-1 amino acid Ile, which is thought to have appeared later in the code.

According to Wong (2005), phase-2 amino acids are difficult to make using prebiotic synthesis and are more readily broken down in natural environments. Wong and Bronskill (1979) showed that Gln and Asn are unstable under heat, even for reasonable conditions (e.g., if all UV photons from the Sun with wavelengths <260 nm arrived at the early Earth and were available for use in prebiotic synthesis and 20 M concentrations of amino acids were reached in 10^9 years within the oceans, then the steady-state concentration of Gln would not have been greater than 3.7×10^{-12} M and the steady-state concentration of Asn would not have been greater than 2.4×10^{-8} M; Wong and Bronskill, 1979). This demonstrates that Gln and Asn were not available from prebiotic synthesis mechanisms and likely originated from coevolving biosynthetic pathways.

1.1.2.2 Premise 2: Pretran Synthesis Function

The second premise, pretran synthesis function involves the ability to synthesize molecules and secondary metabolites such as alkaloids, pigments, and antibiotics and is necessary for the survival and evolution of organisms. 'Inventive biosynthesis' is the term used to describe the development and utilization of novel metabolic pathways, which can be used to synthesize new metabolites and several of the phase-2 amino acids. In the coevolution theory, a subset of inventive biosynthesis pathways, wherein the functional group (moiety) of an amino acid is altered in a given aminoacyl-tRNA by synthesis, resulted in the production of new metabolites or amino acids. These subset pathways are called pretran synthesis. For example, in pretran synthesis, a Met-tRNA is transformed into formyl-Met-tRNA (Wong 2005:Table 2). Pretran synthesis is advantageous because the novel amino acids are generated already pre-attached to precursor tRNAs. In regular inventive biosynthesis, newly synthesized amino acids first must locate an appropriate tRNA to be encoded.

When coevolution theory was first published, pretran synthesis was observed to take place only during the integration of fMet and the integration of Gln and only within Gram-positive bacteria. Pretran synthesis now is known to be widespread throughout all three domains of life and is hypothesized to have been responsible for the integration of Gln, Asn, Sec, and Pyl into protein sequences of many life forms, (Commans and Bock 1999, Krzycki 2004, Polycarpo et al. 2004). Among known examples, Asn synthetase and asparaginyl-tRNA synthetase appear to have been derived from a gene for aspartyl-tRNA synthetase. This implies that pretran synthesis of an Asn-tRNA occurred before free Asn amino acids were directly integrated into an asparaginyl-tRNA synthetase (Roy et al. 2003, Francklyn 2003). Other evidence for the prevalent use and evolutionary

significance of pretran synthesis is the fact that pretran synthesis pathways that utilize aminoacyl-tRNA as their substrates can be found in porphyrin synthesis pathways, cell wall peptide synthesis pathways, N-modifications to proteins, and the ubiquitin pathway (Danchin 1989, Ibba et al. 1997).

1.1.2.3 Premise 3: Biosynthetic Imprints on Codon Assignments

The third premise, biosynthetic imprints on codon assignments, was considered from two perspectives. The first perspective (Amirnovan 1997) found that there was a 0.001 random probability for codon allocation and biosynthetic pathways being correlated. This was a greater probability than the 0.0002 result produced by Wong (1975). Removal of two of the precursor-product pairs (Amirnovan 1997) from the reassessment increased the probability to 0.036. These results suggest that the relations between codon location and biosynthetic pathways are statistical only and have no real-world significance. Wong counters this by citing two studies (Di Giulio 1999, Di Giulio and Medugno 2000) that show that the removal (Amirnovan 1997) is problematic and not easy to validate.

The second perspective (Ronneberg et al. 2000) also involved alterations to the biosynthetic relationships described by coevolution theory and was able to produce a significant probability of 0.0062. The number of sense codons was decreased from 61 to 45, severely impacting the number of possible alternative genetic codes. When combined with the alterations made to methods for the evaluation of probabilities, all possible significant correlations between biosynthetic pathways and codon allocations were eliminated. Again, Wong (2005) cited evidence (Di Giulio 2001) to contest some of the

basic postulates in producing the 0.0062 value (Ronneberg et al. 2000). For example, in producing that value, researchers proposed that the same enzyme that catalyzes the transformation of homoserine to Thr also is responsible for the reverse reaction (Ronneberg et al. 2000). However, real-world cell metabolism mechanisms usually make use of different enzymes for forward and reverse interconversion reactions that are energy-dependent (e.g., phosphofructokinase is the enzyme responsible for converting fructose-6-phosphate into fructose-1, 6-biphosphate, but a different enzyme called fructose-1, 6-biphosphatase, carries out the reverse reaction). Wong (2005) argued that, if the assumption that the same enzymes are responsible for catalyzing forward and reverse reactions, then a significant number of possible metabolic conversions are obstructed, including gluconeogenesis and Gln hydrolysis. The dispersal of glutaminyl-tRNA synthetase and asparaginyl -tRNA synthetase species also dispels the notion that pretran synthesis might not be a remnant of biosynthetic pathway evolution. The absence of glutaminyl-tRNA synthetase and asparaginyl-tRNA synthetase and dependence of free Gln and Asn being placed into protein sequences by pretran synthesis are observations deeply significant to the build-up of amino acids in a PGC (Siatecka et al. 1998, Tumbula et al. 1999, Xue et al. 2003).

Wong (2005: Figure 1) displayed the SGC with the amino acids colour-coded according to their biosynthetic families. The Asp amino acid family extends over the left-centre row of the ANN codons. Amino acids that were synthesized from Ser or Glu possess row and column similarities. Row and column similarities are also observed between the precursor-product pairs Val-Leu and Phe-Tyr, indicating that biosynthetic imprints can be extrapolated from the SGC using statistical analysis or examination.

Codon blocks have an elevated amount of sibling amino acids. Only 10 (4.3%) of the 231 possible pairings of 20 amino acids in the SGC, based on the biosynthetic relationships postulated in coevolution theory, are sibling pairs. However, of these 10 amino acid sibling pairs, only 5 pairs share a 4-codon block with each other: Ile-Met, Asn-Lys, Cys-Trp, Cys-Sec, and Trp-Sec. Wong (2005) proposed that this is evidence for enrichment among these codon blocks being 10-fold greater than in the others. Of particular note is the codon block containing Cys, Sec, a stop codon, and Trp. When compared to all other codon blocks, this block is most dissimilar in terms of physical properties; however, Cys and Trp being a sibling amino acid pair readily explains this anomaly, even when error minimization fails to do so.

Pretran synthesis was hypothesized as the key process by which biosynthetic imprints are left on the genetic code. The basic premise of pretran synthesis is that an acceptor tRNA of a precursor amino acid is obtained by the product amino acid. The product amino acid then is assigned to codons that previously belonged to the precursor by the acceptor tRNA, granting the product amino acid a neighboring location among codons of the precursor. Examples of this are seen in Gln being assigned to codons that are hypothesized to have been assigned previously to Glu; Asn being assigned codons that were assigned previously to Asp, and Sec, Cys, and Trp being assigned codons that belonged previously to Ser. Wong (2005) also suggested that pretran synthesis of cystathionine-tRNA and homocysteine-tRNA was the source for Met through mechanisms involving homoserine and Thr, with homocysteine, homoserine, and cystathionine possibly being encoded temporarily amino acids before Thr and Met were included. Wong (2005) additionally argued that pretran synthesis involving a Glu-5-

semialdehyde-tRNA could be employed to create a Pro-tRNA from a Glu-tRNA. Another position in the SGC that could be influenced by pretran synthesis is the transformation of Glu into Arg, with L-glutamate being used to form ornithine; ornithine being transformed to citrulline; and citrulline finally being converted to Arg. In this case, ornithine and citrulline could have been amino acids encoded temporarily, similar to homocysteine, homoserine, and cystathionine (Jukes 1973).

As has been stated previously, a greater structural similarity between amino acids of precursor-product pairs allows for more chances to obtain precursor codons via greater competition when binding to precursor isoacceptor tRNAs. For example, Val would contend for codons with Leu once Leu had been synthesized. In another comparable scenario, Tyr and Phe share similar chemical structures (in addition to the fact that Tyr can be formed from Phe through pretran synthesis), meaning that inventive biosynthesis alone may provide an adequate explanation for neighbouring codons being shared by Tyr and Phe. In some cases, the attachment of a product amino acid to a precursor acceptor tRNA actually is favoured by metabolic channelling. Situations like this include products whose biosynthesis is reliant on an intermediate molecule that possesses thermolabile properties. Examples of such intermediates include acetyl-Glu-semialdehyde, when Arg is synthesized from Glu, and alpha-ketobutyrate, when Ile is synthesized from Thr (Wachtershauser 1988, Edwards 1996).

In hyperthermophilic environments, thermolabile intermediates must be buffered from breaking down so that they may participate in future reactions (Moser et al. 2001, Manssant et al. 2002). In a pre-LUCA (Last Universal Common Ancestor) world, many environments where early evolution is hypothesized have to occurred would have

favoured hyperthermophiles (e.g., hydrothermal vents), which would have favoured preservation of thermolabile intermediate molecules through metabolic channelling (Xue et al. 2003). Furthermore, evolutionary advanced tRNA-independent reactions eventually could supplant tRNA-dependent pretran synthesis. For example, in animals, fungi, and some bacteria, 5-aminolevulinic acid synthase condenses succinyl-CoA and Gly in a single step, a process that has replaced pretran synthesis of Glu-1-semialdehyde to make 5-aminolevulinic acid and tetrapyrrole compounds (Moser et al. 2001). It even is possible that certain pretran synthesis mechanisms vanished prior to the arrival of the LUCA; therefore, such mechanisms would be absent in any surviving extant organisms. All of these processes affect how product amino acids would have been assigned codons of their precursors, thus leaving an imprint of biosynthetic pathways on the SGC, itself.

1.1.2.4 Premise 4: Amino Acid Mutability

The fourth premise, amino acid mutability, involves how labile already encoded amino acids could have been. When the coevolution theory was published, one of the properties that it predicted was the ability of the evolving genetic code to be mutable. No evidence for this kind of mutability within the code is known. Amino acid analogues exist and can, over short time periods, be integrated into a protein to the point that even growth yield is enhanced (Pezo et al. 2004). With respect to this information, the only way to assess mutability was to perform alterations on the SGC using the gram-positive bacteria *Bacillus subtilis*. In experiments, mutants had their genetic codes modified such that Trp was replaced with 4-fluorTrp, and the analogues supported continuous cell growth (Wong 1983, Bronskill and Wong 1988). Trp, alone, became an inhibitory amino acid analogue and was unable to support continuous growth. In addition to *Bacillus*

subtilis, *Escherichia coli* and bacteriophages were able to undergo genetic adaptation for growth on 4-fluorTrp (Bacher and Ellington, 2001, Bacher et al. 2003, Bacher et al. 2004).

These types of mutations have profound effects on the fundamental arrangement of proteomes, as encoded amino acids are mutated, which imparts effects on protein sequences. Such demonstrations of the mutability of encoded amino acids in the SGC lend credence the idea of directing the evolution of the current SGC. Using a 'bottom-up' approach to mutate the SGC, orthogonal tRNA - aminoacyl-tRNA synthetase pairs that utilize suppressor tRNAs (and do not participate in cross reactions between the tRNAs and aminoacyl-tRNA synthetases of other organisms from different domains of life) provide one method of directing further evolution of the SGC. Using a 'top-down' approach to mutate the code involves mutations to organisms themselves (Bacher and Ellington 2003, Kowal et al. 2001, Magliery et al. 2003, Kwok and Wong 1980). Mutability of the genetic code also allows for the exploration of new hypotheses, like construction of novel proteins and the possibility that the SGC will evolve into a new stage with an expanded amino acid repertoire beyond the 20 standard amino acids (Bock 2001). Wong (2005) even suggested that the SGC will resume its slow encoding of more amino acids in spite of a 2-3 billion year period of inactivity.

After reanalyzing the 4 main premises of the Coevolution theory, Wong (2005) discussed the universality of the SGC. The fact that it underwent a great expansion and then became dormant indicates a dynamic process of evolution during the early phase of its assembly. According to coevolution theory, if the evolution of the PGC started with either 1 or 20 amino acids and 3 stop codons, then there are $N = 2 \times 10^{19}$ possible

alternate genetic codes. This number can be reduced greatly if it is assumed that the PGC began with only 5-10 founding amino acids and used subdivision of codon blocks to arrive at the final set of 20. In this case, only within the separate founder domains would permutations of codon position occur, causing N to decrease by a factor of 10^{11} (i.e., the total number of possible alternate genetic codes to drop to 2×10^{8}; Wong 1988). For the SGC to become prevalent among organisms requires a massive decline in N. The number of codes that must have been eliminated and the available time within which the LUCA could have done so raises some issues. For example, given 2×10^{19} possible alternative genetic codes, a LUCA possessing the SGC competing against other organisms possessing other codes, and one alternative genetic code eliminated once every second, 2 billion years still would be insufficient to have removed all possible alternative genetic codes.

Covalent bonding between aminoacyl cofactors, peptidyl cofactors, aminoacyl-oligonucleotides, peptidyl-oligonucleotides, free amino acids, or free peptide sequences to early ribozymes can be driven by enhanced catalytic activity that may have begun evolving in an RNA World (Seligmann and Amzallag 2002, Szathmary 1999, Wong and Xue 2002, and Wong 1991). RNA eventually abandoned its catalytic activities in favour of encoding peptide sequences, an event that was facilitated by the growth of peptide functional groups (moieties) on early ribozymes. The catalytic potential and activities of peptide moieties on ribozymes eventually outstripped those of RNA backbones as they grew in length. The aminoacylation of tRNA by aminoacyl-tRNA synthetase can take place in sites other than terminal 3'-OH groups, providing evidence on how precursor molecules that linked ribozymes with peptides or amino acids could have evolved into

specialized aminoacyl-tRNA synthetases (Ibba and Francklyn 2004). These newly evolved aminoacyl-tRNA synthetases would have linked amino acids only to tRNA 3' terminal groups. Linkage of peptides and amino acids to sites on tRNA other than the 3'-OH group also may indicate that protein synthesis initially would have required no aminoacylation of a 3'-site specifically on the tRNA.

Wong (2005) proposed that stereochemical interaction between tRNA anticodons or codons and their respective amino acids was key to the expansion of the PGC when the phase-1 amino acids were being incorporated (Pelc 1965, Dunnill 1966, Melcher 1974, Hendry et al. 1981, Shimizu 1982, Yarus 1988, Knight et al. 2003, Seligmann and Amzallag 2002). This would have been prior to any biosynthetic pathways evolving and synthesis of phase-2 amino acids. Stereochemical interactions also could have guided codon assignments to amino acids through interactions of tRNA that possessed complementary anticodons, allowing the coding system to take on novel tRNA species (Guimaraes and Moreira 2004).

The expansion of codons also was suggested by Wong (2005) to have been the method through which codons were added gradually in stages to steadily build up the SGC (citing research by Eigen and Schuster 1978, Jurka and Smith 1987, Delarue 1995, Davis 1999, Yang 2004, and Guimaraes and Moreira 2004). However, this idea implies that not all of the codons would have been available immediately for amino acids to use and many codons would have been more favourable than would have others, as some amino acids were added to the PGC. GNN codons were most likely among the first codons to be preferred (Wong 1981, Trifonov, 2004) due to their thermostability and pair complementarity (e.g. GGC and GCC). Wong (2005) suggested that proper modifications

to the nucleosides located within tRNA anticodon-loops were necessary for equalization of H-bond strengths of differing codon-anticodon pairs. This eventually would have allowed the utilization of the entire repertoire of the 61 sense codons. It would not be until phase-2 of development (e.g. the time in which the phase-2 amino acids were added) that the process of amino acid biosynthesis and biosynthetic pathways predicted by coevolution theory would become significant factors.

Another indicator for the universality of the SGC is the observation that anticodons of tRNAs are recognized as identity elements by most of their cognate aminoacyl-tRNA synthetases, with exceptions being serinyl-tRNA synthetase and alanyl-tRNA synthetase (Giege et al. 1998). This pattern is observed using *Bacillus subtilis* as a bacteria model, wherein the tryptophanyl-tRNA synthetase enzyme recognizes the anticodon and 3'-discriminator base of the cognate tRNA as identity elements, which can influence aminoacylation of tRNA (Xue et al. 1993, Guo et al. 2002). The amino acids Asp, Glu, Phe, Ser, and Val have codon domains that are row-centered and column-centered, indicating a division between the phase-1 amino acid families, which may have been caused by ancient aminoacyl-tRNA synthetase anticodon identity elements. For example, the NNU class of anticodons may have been preferred by ancient aspartyl-tRNA synthetases, which in turn would have caused the Asp family to be assigned the ANN class of codons. Similarly, the NAN class of anticodons may have been preferred by ancient valyl-tRNA synthetases, leading to the Val family being assigned the NUN class of codons.

A binding site for an amino acid substrate on aminoacyl-tRNA synthetase and an anticodon binding site are necessary to utilize an anticodon as an identity element.

Therefore, in terms of associating phase-1 amino acids with codons and anticodons based on stereochemical properties, the use of anticodon identity elements provides an alternative mechanism to stereochemical interactions. Overall, the two binding sites on the aminoacyl-tRNA synthetase possess stereochemical qualities that could have guided specific types of amino acids to become part of the evolving genetic code via specific anticodons and codons. The growth and expansion of the PGC necessitated new amino acids being activated via the enlistment of new aminoacyl-tRNA synthetases in addition to prior aminoacyl-tRNA synthetases for older amino acids becoming adapted to new amino acids.

These changes may be reflected in the SGC where strong similarity is seen between tyrosyl-tRNA synthetase and tryptophanyl-tRNA synthetase, as well as the three-way similarity observed between valyl-tRNA synthetase, leucyl-tRNA synthetase, and isoleucyl-tRNA synthetase (Nagel and Doolittle 1995). Class I and Class II aminoacyl-tRNA synthetases dock to the terminal 2'-OH and terminal 3'-OH, respectively, on tRNA, which allows recognition of disparate identity elements that are bound to similar tRNAs. This might have allowed aminoacyl-tRNA synthetases to have recruited novel amino acids into the PGC. This dual docking mechanism might explain why two different classes of aminoacyl-tRNA synthetases exist. Research (Ribas and Schimmel 2001) supports the first premise of coevolution theory (necessity of amino acid biosynthesis) by suggesting that the pairing of ancestral aminoacyl-tRNA synthetases corresponded to fewer amino acids and tRNA identities in the PGC. This accord with the arguments made on the basis of the first central premise, that the PGC contained only

phase-1 amino acids that were immediately available from prebiotic mechanisms, while other amino acids were added later through biosynthetic pathways.

On the basis of the previously discussed ratio amino acid biosynthesis: error minimization: stereochemical interaction 40000000:400:1, Wong (2005) argued that the prevailing force in building the SGC was the evolution of biosynthesis pathways. As more amino acids were incorporated into the growing PGC, the diversity and functionality of proteins would have increased dramatically and resulted in diffusion-controlled reaction kinetics that improved catalysis rates in enzymes (Wong, 1975). Once all 20 standard amino acids had become encoded, incorporating additional amino acids with their own codon assignments would have caused disruptions at sites within proteomes. At that point, the expansion of the genetic code essentially ended, with most major protein alterations and phase-III amino acid incorporation becoming superseded by post-translational modification processes (Wong 1976). The phase-III amino acids, such as Hydroxyproline, Cystein-heme, adenylyl-Tyrosine, Histidine-flavin, pyro-Glutamic acid, and glycosyl-Asparagine were incorporated into proteins and added to the amino acid repertoire without direct genetic encoding through post-translational modification. In this way, the advantages conferred by placing the phase-III amino acids into limited numbers of protein residues are greater than the slight disruptions they cause by their limited presence in the proteome.

Wong (2005) also briefly discussed the role that coevolution theory can play in identifying the root of the tree of life. According to coevolution theory, early tRNA may have been enriched as paralogs (e.g., single tRNA species may have diverged after replication events into different tRNAs with new functions). Phase-II amino acids that

obtained isoacceptor tRNA from their precursor amino acids through pretran synthesis mechanisms would have led to the formation of new tRNA paralogs. In analyses of 60 different Archaea, Bacteria, and Eukarya genes, a high degree of sequence similarity for tRNAs that can be charged with different amino acids (termed "alloacceptors") was identified in hyperthermophile Archaeans (Wong 2005:Figure 3, Wheeler et al. 2004, Altschul et al. 1997, Tumbula et al. 1999, Xue et al. 2003). This observation can be taken as evidence that several tRNAs with similar sequences were subjected to duplication, coevolution with amino acid biosynthesis pathways, and evolutionary alterations that scattered the spacing of their sequences. Ultimately, this would be the root of evolution for the PGC.

The amount of clustering of tRNA within genomes provides a useful means for locating the most likely position of the LUCA on a tree of life. This is done by measuring the distance, D_{allo}, to an alloacceptor tRNA. A pair of tRNAs that have been enriched and possess a $D_{allo} < 0.2$, compared with other amino acid pairs that are biosynthetically related, provides evidence supporting coevolution theory (Xue et al. 2003). An archaean, *Methanopyrus kandleri*, requires pretran synthesis to place glutamine and asparagine into its proteins, as it does not possess a glutaminyl-tRNA synthetase or an asparaginyl-tRNA synthetase. This indicates that the Gln-tRNA and Asn-tRNA found in *Methanopyrus kandleri* resulted from pretran synthesis during the evolution of the PGC, not prebiotic synthesis.

Measurement of D_{allo} and observations of the presence and absence of certain aminoacyl-tRNA synthetases place the LUCA closest to *Methanopyrus kandleri* on most trees of life. Identifying the LUCA on a tree of life is important to the coevolution theory

because the identification would aid in explaining the near universality of the SGC. If life originated multiple times, the near universality of the SGC would be problematic.

1.2 The Triplet Code from First Principles (Trifonov, 2004)

Trifonov (2004) theorized on the temporal order of appearance of amino acids and their respective codons. The related analysis was performed under the assumption that only a subset among the 64 codons were immediately available to the PGC, and others appeared over time in tandem with their assigned amino acids. Using 60 different factors, Trifonov (2004) constructed a chronology of amino acids that remains generally consistent with current knowledge. Trifonov (2004) proposed the consensus order: Gly, Ala, Asp, Val, Pro, Ser, Glu, (Leu, Thr), Arg, (Ile, Gln, Asn), His, Lys, Cys, Phe, Tyr, Met, Trp. Remarkably, the amino acids found in permutations of the experiments by Miller (1953) appear first, while codon capture-associated amino acids appear last. Based on the consensus chronology, Trifonov (2004) proposed a chronology for the appearance of triplet codons as well. The codon chronology presented in his study is consistent with the rules of thermostability, complementarity, and processivity. The overall reconstruction of SGC evolution followed five fundamental premises: 1) abiotic beginning, 2) complementarity, 3) thermostability, 4) processivity, and 5) codon capture. The latter 4 are described subsequently.

1.2.1 Complementarity

The complementarity & thermostability premises were suggested initially by Eigen and Schuster (1978), utilizing hypercycles as the operative mechanism. The complementarity premise comes into effect in considering the early stages of

oligonucleotide formation and RNA self-replication and has far-reaching effects on the order in which codons would have appeared and their assignments to amino acids.

The ability to self-replicate was requisite to the origin of life, as it would have provided a way for information to be passed from one generation to the next. Eigen and Schuster (1977) explored a model for RNA self-replication involving hypercycles, where the products of prior reactions in one cycle can act as catalysts for the precursors of the next cycle (Eigen and Schuster 1977: Figure 7). Mathematical modeling showed that hypercycles linked like this can result in the formation of quasispecies, which are unstable groups of RNA genotypes characterized by abnormally high mutation rates (Biebricher and Eigen 2006). In RNA quasispecies, offspring contain mutations relative to their parental strands. Below a certain critical temperature, two complementary RNA strands join together in a complete or partial duplex structure. In a temperature driven model, a four-step reaction occurs:

1. the 4 nucleotides present in the environment are ligated into different length polymers;
2. the polymers associate into duplex structures (i.e., each polymer binds to its complementary strand) below a certain critical temperature when complementary sequences encounter each other;
3. non-hybridized sites on RNA are broken down into individual nucleotides;
4. as more nucleotides are released back to the environment, the chance that they will be ligated together again is great, and the cycle begins anew.

Such hypercycles ensure that strands that have complementary components are preserved due to the greater structural durability of a double-helix versus a single-helix (Eigen and Schuster 1977:Figure 11) . In this way, nucleotide sequences that pair complementary sequences on other strands are more likely to survive in an environment and be assigned to an amino acid and integrated into protein sequences. The complementarity rule would

have acted on genetic material undergoing hypercycles in the early stages of prebiotic chemistry, when oligonucleotides were being synthesized. One drawback of the hypercycle theory is that it requires nucleotides, which are relatively complex biomolecules that generally do not form in prebiotic synthesis experiments (Eigen and Schuster 1977). Johnston et al. (2001) demonstrated that short length, artificially produced ribozymes could synthesize shorter strands of RNA on their own. A 189 base-pair RNA was capable of producing small RNAs that were 11 nucleotides in length and adding them to a primer. The 189 base-pair RNA also was highly accurate in terms of its polymerization ability. When primers were extended by the RNA with 11 additional nucleotides and then cloned, 1088 of 1100 nucleotides of the primer sequence matched the template sequence. These studies show that RNA has the capacity to perform self-replicating functions, which would have been a necessary feature of RNA enzymes (ribozymes) for the origin of life. This study provided experimental evidence that self-replication of RNA was possible with a template-lengthening ribozyme.

1.2.2 Thermostability

The thermostability premise is relatively simple, as it decrees that codons with the greatest thermostabilities (e.g., those with the highest melting enthalpy values in kcal/mol) will be the first to appear in the genetic code with amino acid assignments. Generally, those codons containing nucleotides that possess three hydrogen bonds as opposed to two will be much more favourable early in the PGC due to their thermostability. For example, sequences or codons with high glycine and cytosine content will be available more readily for taking on amino acid assignments because their codon-anticodon interactions with complementary nucleotides will be more stable.

Stability of these codon-anticodon interactions would have been vital, as no other outside forces yet would have been available for additional stabilization of such simple molecular systems. This factor would have been especially important in an early Earth environment, where temperatures are thought to have been much higher than today (Trifonov 2000).

1.2.3 Processivity

The processivity premise establishes that new codons in the PGC would have arisen through progressive point mutations of previously existing codons rather than *de novo* synthesis. In other words, new codons were derived from chronologically earlier ones. When the processivity rule is combined with the complementarity rule and the thermostability rule, newly added codons may be argued to have arisen as derivatives of chronologically prior codons, as complementary pairs, and in decreasing order of thermostability. Combining premises allows a consensus chronology of codons to be constructed consistently, even when additional data of amino acid chronology is taken into account.

The amino acid chronology presented in Trifonov (2004) is based on 60 different factors. Some of these factors are interrelated while others are independent; therefore, they were split into two different, non-mixable classes. Interrelated factors were combined and represented by a single, averaged ranking for the entire group. The independent factors were sorted into a 'Single-Factor Criteria' class while all interrelated factors were sorted into a 'Multi-Factor Criteria' class. A chronological order of appearance for all 20 amino acids then was calculated once for each criterion. A positive correlation between the two amino acid chronologies was interpreted as reflecting a

common trend, as the Single-Factor criteria should be independent. The Multi-Factor criteria consist of assorted synthetic, interrelated hypotheses about the SGC foundation and development. These hypotheses involve many of the same basic ideas but with different emphasis depending on the topical knowledge of the author, essentially making the Multi-Factor Criteria individual, subjective, biased opinions.

Concerning the Single-Factor Criteria specifically, the interrelated factors were combined into groups and given a single, average ranking that represented the group as a whole. Ultimately, 17 independent rankings for Single-Factor Criteria were obtained, with the remaining criteria being excluded from the consensus order calculation if they did not show a positive correlation with the other factors. (Trifonov 2004:Table 1 and 2). On the basis of chronologies derived from correlations, two characteristics can be deduced: 1) the initial 10 abiotically synthesized amino acids (e.g., those made through prebiotic synthesis mechanisms using chemical materials from environments) take chronological priority over the rest of the amino acids and 2) amino acids whose respective codons have (or seem to have) been taken from the codon repertoires of other amino acids always appear at the end of the Single-Factor Criteria chronology. Trifonov (2004) conceded that knowing which amino acids borrowed codons from other amino acids cannot be deduced with absolute certainty using only *a priori* knowledge.

Multi-Factor Criteria differ in that most of them are interrelated but none are placed into groups. Instead, 13 independent hypotheses and theories about the evolution and origin of the genetic code were considered, each with their own amino acid chronology. Many of the Multi-Factor Criteria are interrelated due to the fact that many of them involve identical concepts or ideas. For example, coevolution of codons and

amino acids (Dillon 1978, Wong 1976, 1981, and 1988, and Wachtershauser 1988) are based on at least two clear notions. For instance, analysis of codon-anticodon interactions (mRNA-tRNA) and amino acids with larger codon repertoires being the earliest in the genetic code are the two principles utilized by the chronology suggested by Jukes (1973) chronology (Trifonov 2004:Table 3 and 4). The same two characteristics that were observed in the Single-Factor amino acid chronology can be observed in the Multi-Factor amino acid chronology: 1) the initial 10 amino acids are those that can be formed from prebiotic synthesis and 2) amino acids that appear to have borrowed codons from previously established repertoires appear at the end.

Based on the analyses, a final consensus chronology for amino acid appearance was devised. Amino acid rankings in both analyses were consistent. While Trifonov (2004) acknowledged that the order of amino acids for both Single- and Multi-Factor Criteria is speculative, he stated that it still warrants significant consideration because all past and current theories and ideas are deliberated when deriving consensus chronologies. The fact that the amino acids found in Miller-type experiments appear first and amino acids that performed codon capture appear last suggests that the consensus has merit. (Trifonov 2004: Table 5). Once a final consensus chronology for amino acids has been postulated, a temporal order of associated codons may be derived.

The codon chronology, at its core, follows the three rules of complementarity, thermostability, and processivity. Trifonov (2004) stressed that these three premises can be applied simultaneously to all triplet codons without conflicting with one another. Based on the complementarity and thermostability premises inferred by Eigen and Schuster (1978), GCC and GGC are the most thermostable complementary codon pair

and likely corresponded to the earliest amino acids of the chronology, glycine and alanine. This is supported by the fact that glycine and alanine have the greatest yields in Miller-type synthesis experiments. Interestingly, this seems to indicate that glycine and alanine appeared simultaneously. The next amino acids to appear in the consensus chronology are valine and aspartic acid; therefore, the next codon pair was assigned to those two. The modern triplet codons GUC and GAC are most strongly associated with valine and aspartic acid, respectively. They also have the next-highest melting enthalpy of all complementary codon pairs in addition to being complementary to each other. GUC and GAC notably can be derived from GCC and GGC through point mutations at the second position. A transition mutation (purine-to-purine or pyrimidine-to-pyrimidine) would have been the energetically 'cheapest' route, with GGC being changed to GAC (or GCC being changed to GUC), after which complementary strand copying of GAC would have yielded GUC (or GAC).

As with the case of glycine and alanine, the concurrent appearance of the GUC and GAC codons suggests that valine and alanine also were introduced into the PGC at the same time. From there, proline should have been the next amino acid to appear based on the basis of the consensus chronology. A transition mutation (energetically less costly) could have occurred at the third position in the GGC codon to produce GGG. Complementary copying of this triplet would have yielded the CCC codon, which in modern day genetic codes is the most-frequently used of the 4 codons that encode proline. Trifonov (2004) proposed that the rest of the codons ultimately assigned in the SGC were derived this way, with mutations at the third position of an earlier codon and subsequent complementary copying of that codon.

Overall, the reconstruction of codon chronology from this perspective results from mutations at third positions of assigned codons followed by complementary copying to form new codons. These new codons also happen to be the most-likely candidates for the next amino acid to appear in accordance with the amino acid consensus chronology. Trifonov (2004:Figure 1) provided an illustration of the complete codon chronology as aligned with the amino acid consensus chronology and three premises of thermostability, complementarity, and processivity. The order of appearance of amino acids in that illustration is not exactly the same as their order in the amino acid consensus chronology. This is due to amino acid rankings having a margin of error of ±0.7, which causes some of the neighbouring amino acids (such as leucine and threonine) to be given ranks that are problematic to discern to a precise degree. As a result, the order of these closely ranked amino acids may be adjusted locally without negatively affecting the overall consensus chronology. For the restoration of the codon chronology to be completely coherent with the thermostability premise, the local orders (T, L) and (N, I, Q) must be changed to (L, T) and (I, Q, N), respectively. The difference in ranking between T and L is ±0.5, whereas the difference in ranking between N, Q, and I is ±0.1.

The codons ACC, CGC, and AGC occur after the 5th step in the codon chronology and are assigned to threonine, arginine, and serine, respectively. These three codons use, as their complementary templates the codons of glycine (GGU) and alanine (GCG, GCU). The thermostability premise suggests that the UGC codon was formed next in the chronology. However, today this codon encodes cysteine, which had not yet appeared based on the amino acid consensus chronology. Therefore, Trifonov (2004) suggested that UGC, UGG, and UGU originally functioned as stop codons. The advantage to having

stop codons appear at this stage is likely two-fold: 1) synthesizing a single protein or peptide would have required translation of an entire strand, something that would have been energetically inefficient, even when considering the self-catalyzing abilities of RNA replication; possession of termination codons would have allowed discrete sequences to be translated, allowing for vastly greater efficiency during translation processes; 2) no cognate tRNAs for termination codons exist in the SGC; instead, the termination codons bind to eukaryotic or prokaryotic protein release factors and GTP; these release factors mimic tRNAs by recognizing an in-frame stop codon and causing hydrolysis of peptide chain ester bonds with the tRNA peptidyl site; this reaction is catalyzed by the ribosome peptidyl transferase region; the end result is the dissociation of the ribosome from the mRNA (Kisselev, Ehrenberg, and Frolova 2003). In the PGC, when not all amino acids had yet been encoded, formation of such release factors from the limited amino acid repertoire would have been unlikely; therefore, the stop codons presumably functioned only after the initial 8-10 amino acids had been incorporated into the PGC, so that proteins with greater structural and functional complexity could be made (e.g., the release factors that are necessary to terminate translation).

In later steps (Trifonov 2004:Figure 1, steps 10-18), the codons that appear in the consensus order are derivations of those that already are assigned to amino acids. Complementary codon pairs in these stages are organized such that they appear in order of decreasing thermostability. The next step (Trifonov 2004:Figure 1, step 19) involved the introduction of isoleucine to the genetic code, assigned the codon AUC. The AUC codon, itself, is a complementary derivation of GAU for aspartic acid. The amino acid consensus chronology indicates that glutamine appeared after isoleucine and derived its

codon from complementary copying of GUG to CAC. However, CAC currently is assigned to histidine in the SGC, so it may have been assigned originally to the glutamine repertoire and later captured by histidine. Subsequent stages (Trifonov 2004: Figure 1, steps 21-32) display the remainder of the codon vacancies being filled. Some instances of codons becoming dissociated from their original assignments and assigned to other amino acids then ensue. For example, codons UUC and UAC originally encoded leucine and a stop codon, respectively, but later were reassigned to phenylalanine and tyrosine. A comparable scenario occurs with codons UUU and AAA, which originally encoded leucine and asparagine, respectively, but later were captured by phenylalanine and lysine. UUU and AAA are the least thermostable among the 64 codons and are the final complementary codon pair. Amino acids in the PGC that appeared after the initial assignment of UUU and AAA captured their codons from previously existing repertoires.

The codon capture phase represents the final stage in the codon chronology (Trifonov 2004). The time period marking the introduction of asparagine and histidine is inferred from the succession of amino acids in the consensus chronology, without constraint from the thermostability rule. According to Trifonov (2004), the original set of termination codons (UGN) would have facilitated a UGG to tryptophan conversion. This hypothesis is supported by a study from Yamao et al. (1985), who showed that individuals in the bacterium species *Mycoplasma capricolum* utilize UGG and UGA codons to encode tryptophan. Capture of codons form the termination repertoire also may have been involved in the conversion for cysteine as well, currently encoded by UGC and UGU, but also encoded by the UGA stop codon in the species *Euplotes octacarinatus* (Meyer et al. 1991). The arrival of tyrosine to the genetic code also resulted in the capture

of codons from the second set of UAN terminator codons, specifically UAC and UAU. The AUG codon most likely originally belonged to the isoleucine block (AUN) but was captured later by methionine. In yeast, methionine likely captured AUA in addition to AUG (Barrell et al. 1980). Contrarily, no instances of codon capture for UUY by phenylalanine from the leucine UUN repertoire are known, nor are there any instances of codon capture for CAY by histidine from the glutamine CAN repertoire. Trifonov (2004) also noted that the CAN codon block underwent a schism into CAR for glutamine and CAY for histidine, consistent with Wong's (1976) proposal that the introduction of histidine to the CAN block caused the split. Additionally, two recent amino acids absent from the amino acid consensus chronology illustration were incorporated into the PGC: selenocysteine and pyrrolysine (Chambers et al. 1986, Srinivasan et al. 2002). The UGA codon is hypothesized to have been secured by selenocysteine from the existing UGN repertoire, leaving no codons from the UGN group unassigned to an amino acid. Pyrrolysine codon capture behaved in a manner similar to tyrosine, with the UAG codon being usurped from the UAN codon group. Theoretically, additional amino acids could have been added to the PGC so that one codon was assigned to one amino acid. However, this is unlikely to have occurred due to the benefits that a moderate amount of degeneracy confers to the translation process.

Trifonov (2004) discussed the value of the amino acid consensus chronology and the reconstruction of the codon chronology. The fact that the codon chronology obeys the three premises, thermostability, complementarity, and processivity, so rigidly and without conflict is interpreted as testament to the quality of the reconstruction. The amino acid consensus chronology, itself, is simplistic and opportunistic at its core: the available

amino acids are assigned codons first, until all codons have been assigned, then new amino acids to the code take their codons from previously assigned repertoires. This pattern is somewhat analogous to patterns in fractal geometry, where the acquisition and repetition of simple axioms (e.g., thermostability, complementarity, and processivity) can result in a seemingly complex system (the triplet code with amino acids assigned to various codons). Another trait of the SGC is the persistent uniformity of the triplet codon system, itself. The triplet codon system that arose billions of years ago is still the same system that is used in the present day SGC, the only major difference being that the early stages of the PGC had a much smaller repertoire of amino acids and codons.

Some of the criteria used for the amino acid chronology are based on contemporary codon properties, somewhat rationalizing the apparent persistence of the SGC. For example, the early steps (Trifonov 2004:Figure 1, steps 1-8) of the codon assignments remain 100% conserved; while latter stages are only mostly conserved, with codon capture events resulting in minor undermining of the conservation. However, this is only the case for the final 7 amino acids in the consensus chronology (histidine, lysine, cysteine, phenylalanine, tyrosine, methionine, and tryptophan) during the codon capture stage. Additionally, Trifonov (2004) reiterated that the entire amino acid consensus chronology and codon chronology reconstruction is speculative, requiring support from data so that this model of SGC origin may be tested and ultimately receive confirmation.

Finally, Trifonov (2004) utilized the reconstructed chronology of the genetic code to explore two ideas: the “glycine clock” and the “binary alphabet.” Reviewing the initial 6 stages of the codon chronology indicated that the first proteins likely relied heavily on glycine for their structures. This may be due to the fact that glycine is structurally the

simplest amino acid and easiest to modify. Glycine, uniquely, behaves as the complement to the other early amino acids (alanine, proline, serine, and threonine) in the first 6 stages. However, later stages of the chronology show that glycine became steadily more diluted, suggesting that the earliest proteins would have glycine as their most abundant constituent. This actually is observed in patches of protein sequences that have not changed since the split between prokaryotic and eukaryotic organisms (Trifonov 1999). Patches that are protein-coding and known to be ancient tend to have higher amounts of glycine in their composition. From this, a "glycine clock" can be formulated, where the older a protein is, the more glycine it would be expected to contain.

The binary alphabet builds on the idea that there may have been more than one PGC during the earliest days of code existence. Since many triplet codons appeared simultaneously due to the complementarity premise, the growing amino acid repertoires could have been divided into two separate groups. Group 1 constituted those amino acids that replaced glycine in early proteins while group 2 constituted those amino acids that replaced alanine. This is possible since complementary strands of RNA would have contained separate codons that were complementary yet assigned to different amino acids. Therefore, it is plausible that the earliest RNA duplexes were composed of a glycine-coding strand and an alanine-coding strand.

These strands would have then been linked to the respective amino acids of their growing alphabet. For example, GGC was first assigned to glycine and was followed by GAC (resulting from a transition mutation at position 2) being assigned to aspartic acid. Afterward, GAG joined the growing strand and was assigned to glutamic acid, itself followed by CGC (complementary to GCG for alanine) that was assigned to arginine.

The group 2 amino acids grew on the complementary strand in a similar manner, with complementary codons being assigned to the amino acids alanine and valine. After this point, all further alterations to codons involved changes only to third position nucleotides (or first positions for complementary codons). The overall result is two different strands that made-up two separate amino acid alphabets with their own respective codon structures. The glycine strand (containing group 1 amino acids) contained codon triplets in the form 'N-purine-N' while the alanine strand (containing group 2 amino acids) contained codon triplets in the form 'N-pyrimidine-N'. As a result of these differences, each strand developed its own amino acid alphabet:

Glycine strand → glycine, aspartic acid, glutamic acid, arginine, serine, glutamine, asparagine, histidine, lysine, cysteine, tyrosine, and tryptophan

Alanine strand → alanine, valine, proline, serine, leucine, threonine, isoleucine, phenylalanine, and methionine

These two strands, together, comprised what Trifonov (2004) referred to as the binary alphabet. Each of the two strands would have encoded "mini"-genes for "mini"-proteins for the two independent strands. The amino acid serine is a commonality between both codes, suggesting a fusion of the alphabets after a certain period. Greater lengths of sequences were formed after the so-called "mini"-genes linked together, resulting in proteins with mosaic structures. These mosaic patterns in proteins should be reflective of the short amino acid patches from the two different amino acid alphabets and still might be present in extant sequences; so, it should be expected that an alternating mosaic pattern from each alphabet is present, albeit in heavily mutated form in the SGC, in

chronologically older proteins. As evidence for this idea, Trifonov (2004) pointed to an autocorrelation analysis of prokaryotic protein sequences performed (Trifonov et al. 2001), which indeed showed that there was an alternation of amino acid alphabets.

Trifonov (2004) proposed that billions of years in the past, this binary code of protein sequences (the alternating alphabets of glycine and alanine strands) would have arisen with the triplet codons. Binary sequences could have encoded these early protein sequence patterns. This type of binary arrangement would have described successfully these protein sequence patterns in their original ancient forms, so long as there were no subsequent mutations that occurred. Given that purine-to-purine or pyrimidine-to-pyrimidine transition mutations were most common due to their lower energetic costs (when compared to transversion mutations), then, despite changes to third or first positions of codons, the N-purine-N and N-pyrimidine-N pattern should have remained conserved. This would have allowed the ancient binary patterns to have retained their conservation for an extended time period. Evidence for this conservation of binary sequences is provided by PAM[61] and BLOSSUM[21] analyses of amino acid replacement matrices, demonstrating that amino acid replacements are confined to each respective alphabet from which they originated and amino acid swaps between the alphabets have been exceptionally rare (Trifonov and Eisenhaber 2004). Arranging these sequences in a binary format reveals an evolutionary significance, as the arrangement reveals relatedness among the established amino acids, especially for those amino acids that are not a part of the SGC.

Trifonov (2004) closed the analyses with possible future tests that could be performed to further examine the origin of the SGC. The earliest possible form of life (as

it is known) that is available for exploration may involve the reconstruction of the triplet codon chronology from first principles. Modeling of this specific phase in SGC evolution is an exciting prospect that has not yet been fully investigated. Trifonov suggested that the four oligomers GCC, GGC, alanine, and glycine constitute a single, multi-component replicator unit. Possible metal catalysts then would support this replicator and encourage self-replication.

1.3 Four-column theory

In a natural follow-up to the triplet code from first principles, Higgs (2009) offered five propositions relating to building-up the SGC in formulating the four-column theory: 1) the first amino acids to be incorporated into a PGC were the ones that could be synthesized most readily under prebiotic conditions, specifically, Gly, Ala, Val, Asp, and Glu; 2) the earliest amino acids were assigned to codons whose first position base was Guanine (GNN), so initial iterations of the PGC involved exclusively codons with first position G; 3) swift evolution of the PGC resulted in a four-column structure, where all codons within each column were assigned to the same amino acid (e.g., NGN = Gly, NUN = Val, NCN = Ala, and NAN = Asp or Glu; 4) further subdivision of columns allowed additional amino acids to be integrated into the PGC, with subsets of codons that had been assigned to early amino acids being reassigned to amino acids that appeared later; 5) amino acids that appeared later were similar in sequence to (i.e., assigned to codon positions previously held by) codons that possessed similar physico-chemical properties, allowing for least possible disruptions to proteins that were already encoded. As a consequence of these factors, amino acid properties of the SGC preserve the four-column pattern inherent in its earliest evolutionary stages.

To determine what type of cost function to use to model incorporation of new amino acids in a PGC, a distance matrix can be constructed to analyze the genetic code based on 8 (later expanded to 9) different amino acid properties (Higgs and Atwood 2005, Higgs 2009): volume from van der Waals radii, bulkiness, polarity, isoelectric point, hydrophobicity scale A, hydrophobicity scale B, surface area accessible to water in an unfolded peptide, fraction of accessible area lost when a protein folds, and polar requirement. Additionally, a weighted distance metric, $d_w(a, b)$=constant x $(\sum_k w_k(z_{ka}-z_{kb})^2)^{1/2}$, was used to assign different weights to different properties. The results revealed that amino acids in columns 1 and 2 are characterized by high degrees of similarity; column 3 amino acids are characterized by moderate degrees of similarity; while column 4 amino acids are characterized by little similarity (Higgs 2009:Figure 1). The clustering is fairly intuitive in terms of amino acid properties: amino acids with the simplest hydrocarbon chains are grouped together, the two acids (Glu and Asp) are grouped together, the two amines (Asn and Gln) are grouped together, the amino acids with the highest pH are grouped together, and amino acids with aromatic side chains are grouped together.

Higgs (2009) measured the fitness and cost of genetic codes that contained less than 20 amino acids by developing a cost function with the weighted distance metric. The general cost function, $\Phi=\sum_i\sum_j F_i p_{ij} g(a_i, a_j)$, was used as a starting point, in which the cost of a given genetic code is determined by calculating the average cost of a substitution error, Φ. The term "g" is the cost of a substitution for an amino acid in a position where a different amino acid usually is placed. This cost is a consequence of reduced protein sequence functionality when non-optimal amino acids are utilized or when misfolding of

proteins due to translation errors causes an increase in toxicity. When the physico-chemical properties of two amino acids are more dissimilar, the value of g increases. Likewise, when the physico-chemical properties of two amino acids or similar, the value of g decreases. Therefore, a low g-value corresponds to a lower substitution cost, indicating the newly substituted amino acid imposes less cost. For these reasons, g can be treated as the distance between amino acids within physical property space. The other terms in the general cost function are: F_i = frequency of a codon *i;* p_{ij} = probability that codon *i* is mistranslated into codon *j*; a_i, a_j = amino acids assigned to codons *i* and *j*, respectively. Other research (Gilis et al. 2001) used $F_i = P(a_i) / n(a_i)$, where $P(a_i)$ is the frequency of the amino acid a_i within a protein sequence and $n(a_i)$ is the number of assigned codons to amino acid a_i. Higgs (2009) used average amino acid frequencies from all three domains of life to avoid bias of all codons in a genetic code having equal frequency for an amino acid.

Following on the work of Freeland and Hurst (1998), a version of the error probability matrix was used to increase the frequency of transition errors and decrease the frequency of transversion errors in addition to making errors at first and third positions in codons more likely than errors at second positions. Higgs (2009) defined a parameter, ε (translational error rates), equal to the probability of codon *i* being mistranslated as codon *j*, so that $p_{ij} = \varepsilon$. This means that p_{ij} and ε are equivalent if codon *i* and codon *j* contain different 3rd position nucleotides or have undergone a transition at position 1. When codons *i* and *j* are distinguished by a first position transversion or a second position transition, then $p_{ij} = 0.5\varepsilon$. If codon *i* and *j* differ by a second position transversion, then $p_{ij} = 0.1\varepsilon$. If codons *i* and *j* differ by more than one nucleotide base mutation, then the

probability of codon *i* being mistranslated as codon *j* is 0 (e.g., 0 ε). The likelihood of correct translation of a codon taking place is thus defined as $p_{ii} = 1-\sum_{i \neq j} p_{ij}$, as errors in first and third positions are more likely than are errors at second positions, so this model seems more accurate than one in which the error rates for individual nucleotides are assumed to be equal.

Many previous analyses had assumed that identical sets of all 20 amino acids were available to randomly shuffled codes. Higgs (2009) modified this by taking into account genetic codes that contained fewer than 20 amino acids available to them, as was likely the case in early Earth conditions when not all amino acid codons had yet evolved. This had an impact on Φ, as the average substitution error cost was calculated over a smaller repertoire of available amino acids. For example, a code with K < 20 amino acids cannot always incorporate the preferred amino acid at all sites because amino acids that ultimately became preferred at some sites might not yet have been incorporated into the evolving code. This, in turn, had an impact on the possible protein sequences that could have been encoded, altering the general cost function to: $\Phi=\sum_{\alpha}\sum_{j}\sum_{i}P(\alpha)\phi_i(\alpha)p_{ij}g(\alpha, a_j)$, where α represents an amino acid among the 20 possible and $\phi_i(\alpha)$ represents how often codon *i* appears at α sites. However, for the K < 20 situation, ε (rate of translation error) tends toward 0, so the probability of a codon *i* being mistranslated as a codon *j* becomes $\delta(i, j)$. This causes the average cost of substitution to approach a limit, from Φ to Φ_0, with $\Phi_{0=}\sum_{\alpha}\sum_{i}P(\alpha)g(\alpha, a_i)\ \delta(a_i, B(\alpha)) / n(a_i)$. In this scenario, Φ_0 now represents the cost of using the best possible amino acid available instead of the preferred amino acid for a given site in a genetic code, not the average cost of translational error.

A reduction in Φ_0 then became the chief criterion for new amino acid additions to the evolving code, not a decrease in average cost of translational errors. Even though the addition of a new amino acid led to a reduction in Φ_0, an increase in ε occurred due to the fact that fewer synonymous mutations of codons occurred when they were subdivided into smaller blocks. The favorability of new amino acid additions to the evolving code thus were determined by the balance of ε and Φ_0. Final manipulations to the cost function were performed to model changes that might have taken place when new amino acids were added to the evolving PGC in a situation where $K < 20$. Higgs (2009) assumed that at any given stage of development, the genetic code was defined both by specific amino acids that had been assigned to certain codons, a_i^{cur}, and by the best available amino acids that exist in the current repertoire, $B^{cur}(\alpha)$. This allowed amino acid frequencies at individual site types to be written as $\phi_i^{cur}(\alpha)= \delta(a_i^{cur}, B^{cur}(\alpha)) / n(a_i^{cur})$ and average cost of substitution errors to be written as $\Phi^{cur}=\sum_\alpha\sum_i\sum_j P(\alpha)\phi_i^{cur}(\alpha)p_{ij}g(\alpha, a_j^{cur})$. If a new amino acid addition to the code resulted in the subdivision of previous codon blocks, then the newly formed genetic code, itself, is described by a_i^{new}, while reassigned codons treated each a_i^{new} as an amino acid that did not exist in the previously evolving PGC. Codons that were not reassigned during the addition of a new amino acid were treated a_i^{new} as a_i^{cur} (i.e., new codes were treated in the same manner as their predecessors).

Instantaneously after a new amino acid had been added, a change in the translation system would have ensued, but gene sequences, themselves, would have remained unchanged. This resulted in codon frequencies remaining unchanged during these intermediate transition states, allowing genetic code costs to be written as $\Phi^{int}=\sum_\alpha\sum_i\sum_j P(\alpha)\phi_i^{cur}(\alpha)p_{ij}g(\alpha, a_j^{new})$. At this point, the cost function involved amino acids

of new, evolving PGCs even though those amino acids were not yet present in their actual codon frequencies. Also, a new set of best available amino acids could be defined by $B^{new}(\alpha)$. If a_i^{new} is equal to $B^{new}(\alpha)$, then codon *i* could be used at site type α. In such a case, the codon frequencies could be calculated as $\phi_i^{new}(\alpha) = \delta(a_i^{new}, B^{new}(\alpha)) / n(a_i^{new})$ and the average cost of a code could be calculated as $\Phi^{new} = \sum_\alpha \sum_i \sum_j P(\alpha) \phi_i^{new}(\alpha) p_{ij} g(\alpha, a_j^{new})$.

The change in cost before and after a new amino acid was added then could be obtained by $\Delta\Phi = \Phi^{new} - \Phi^{cur}$. When $\Delta\Phi$ is < 0, the fitness of the genetic code increased from the addition of a new amino acid. In the early stages of genetic code evolution, adding almost any new amino acid to a given position would have caused an overall fitness increase due to the long term decrease in Φ_0. The change in cost immediately after a genetic code has changed due to new amino acid addition, but before alterations to gene sequences, could be calculated with $\delta\Phi = \Phi^{int} - \Phi^{cur}$. The $\delta\Phi$ is highly correlated to which specific amino acid is added to which specific position in a genetic code. When added to any site without bias, the amino acid will likely cause a positive $\delta\Phi$ value, indicating an overall cost increase. This is because codons that evolved to specify a particular amino acid for a particular site will now specify a different amino acid, destructively interfering with previous protein coding sequences. Natural selection should have precluded such unbiased addition because organisms that house the changed amino acid would have been characterized by lower fitness compared to the rest of the population. An important observation to note is that $\delta\Phi$ has a high likelihood of being positive, despite a negative $\Delta\Phi$ value. This means that additions of new amino acids would have been inhibited by short-term costs, despite the possibility of long-term benefits when gene sequences adapted to new genetic codes.

The GNN code system is proposed to have existed before the four-column code structure had been developed, meaning that all tRNAs would have contained a C at the 3rd anticodon position to complement the G of the 1st codon position. The mutation of C to any of the other three nucleotide, along with the replication of such tRNAs, would have made it easy for a GNN code to build up into a four-column code (Higgs 2009:Figure 2). In extant prokaryotic genomes, at least one tRNA gene is associated with every codon pair (e.g., a tRNA with a G in the wobble position has the ability to pair with either C or U in the 3rd codon position, while a tRNA with a U in the wobble position has the ability to pair with either A or G in the 3rd codon position).

Working under the assumption that each tRNA was able to pair with two codons, a four-column genetic code would have necessitated 8 tRNAs for each amino acid, resulting in 32 individual tRNAs total (Higgs 2009:Figure 3). For example, in the column one assignment of UUN and CUN from Val to another amino acid, an alteration to an aminoacyl tRNA-synthetase must have taken place to change the tRNAs that were being charged. The UUN and CUN codon groups would have corresponded to tRNAs that possessed A and G at the third anticodon position, respectively, while AUN and GUN codons would have corresponded to tRNAs that possessed U and C at the third anticodon position, respectively. Therefore, developing the ability to discern between purine and pyrimidine structures at the third anticodon position would have been of paramount importance for the new aminoacyl tRNA-synthetase. Higgs (2009) assumed that the original aminoacyl tRNA-synthetase of the Val column underwent a duplication event, and the two synthetases diverged and specialized to different codon sets. This gave organisms chances to 'try out' new amino acids in the UUN and CUN codon blocks,

gaining a fitness advantage if the addition of the amino acid was favourable to the functionality of protein-coding sequences. In this scenario, the evolution of a new synthetase ribozyme that could interact with an amino acid that was absent in the previous code was the means by which new amino acids were tried out in an evolving PGC. Amino acids that were added to the evolving PGC in those positions were Leu, Thr, Glu, and Arg, so Higgs (2009) considered the different code costs when those amino acids were added to each of the four possible sites.

Higgs (2009:Figure 4) displayed the cost of an evolving PGC, Φ^{cur}, the cost of the intermediate code immediately after an amino acid had been added, Φ^{int}, and the cost of the new PGC after codons had been assigned or reassigned, Φ^{new}. A selective barrier to adding particular amino acids to specific columns of the PGC existed. The change in cost between Φ^{int} and Φ^{cur} defines the barrier of selection that prohibited a new amino acid from entering the PGC. Leu was the most favourable amino acid to have been incorporated into the first column; Thr was the most favourable amino acid to have been added into the second column; Glu was the most favourable amino acid to have been added into the third column; but there is no favorability toward adding Arg into any column.

Even when the change in cost between the evolving new PGC was negative for all four amino acids, the change in cost between the intermediate code and new code was negative only for Leu. This means that these four amino acids would have conferred long-term benefits onto the PGC, but only the addition of Leu to the code would have resulted in an advantageous codon reassignment for the short-term. For the second column, $\delta\Phi$ is strongly positive for Leu, Glu, and Arg, but mildly negative for Thr,

making it the preferred amino acid in that column. Similarly, Glu has a negative δΦ value for column three, but Leu, Arg, and Thr have positive δΦ values. Column four presents a mild conundrum, as the δΦ is positive for all four amino acids and selection would not have supported the addition of Arg. This is likely due to the fact that Arg has a very different structure from Leu, Thr, and Glu, so adding it to an evolving genetic code at such a stage will cause a higher amount of disruption to protein sequences, despite the fact that protein variety would have increased by the addition. Higgs (2009:Table 2 and 3) showed the δΦ (barriers) and ΔΦ (the net change in cost of a genetic code) values for adding an amino acid to an 8-codon block in the first column when translational error rates are ε=0.05 and ε=0.

The previously described approach was used to determine the likelihood of new amino acids being added to a PGC at particular sites and at specific points in time in PGC evolution. The main factor promoting amino acid addition appears to have been increased diversity and functionality of proteins through an enlarged amino acid repertoire rather than minimization of translational error. However, in the four-column theory, the resultant genetic code is one where minimization of translational error has occurred.

The subdivision of codon blocks that occurred as amino acids were incorporated allowed old codons to be reassigned to previously existing amino acids so that remaining amino acids could be assigned to new codons. By utilizing the cost function and barrier function to measure the likelihood of amino acids being added to certain columns, Higgs (2009:Figure 5) illustrated what the structure of the four-column code may have been after the first 10 amino acids had been incorporated.

1.4 Computational Analysis of the SGC

Degagne (2015) analyzed optimization of the genetic code with respect to 54 amino acid properties. Hypothetical genetic codes (HGCs) for each property were generated by shuffling of amino acid identities within codon blocks and error measure calculations were performed using absolute distance and squared distance metrics. The SGC then was ranked against these HGCs based on its capacity for minimizing physiochemical distances between amino acids whose codons were separated by only a single mutation. Perhaps most notably, results indicated that shuffling stop codon positions imparted no effect on the rank of the SGC, however, SGC rank increased upon the inclusion of stop codons into error measure calculations. A curious observation that arose from this research was that long-range non-bonded energy appeared to yield more optimization than did polar requirement, the previously identified most-conserved property.

A computer program for including stop codons in error measure calculations had been used only once previously (Goodarzi et al. 2004). For error measure calculations, stop codons were treated in the same manner as amino acids but with all stop codon property values set to 0. Depending on the property being analyzed, stop codon inclusion significantly impacted the way that values were allocated among codons. The novel feature of this first iteration of the computer program, AGCT, was that it included stop codons in identity shuffling during HGC generation. The two vertically contiguous stop codons, UAA and UAG, were treated as a single block, while the third stop codon, UGA, was treated as its own block. Though an energetically cheap transition can be done to transform UAA into UGA, prior studies have shown that mutations to second positions

are least likely to occur to maintain a degree of redundancy in the SGC (Haig and Hurst 1991, Alff-Steinberg 1969, Woese 1965).

Simulations involving stop codons were performed under three different modes: 1) stop codon positions fixed with mutations involving stop codons included in error measure calculations; 2) stop codon positions variable with mutations involving stop codons excluded from error measure calculations; and 3) stop codon positions variable with mutations involving stop codons included in error measure calculations. Degagne (2015) elected to exclude the fourth possible combination (stop codon positions fixed and mutations involving stop codons excluded from error measure calculations) since it had been extensively studied in previous literature. Using Welch's t-test as a statistical analysis, noteworthy ranking discrepancies were found between modes 1 (μ=2.10, SD=1.04) and 2 (μ=3.55, SD=1.93), where t_{30}=2.93 and p=0.006, but no significant variations were found between modes 1 and 3 (μ=1.70, MD=0.80), where t_{35}=1.34 and p=0.19. From these data, it was suggested that the deviations between the first and second modes were mainly the result of including or excluding the stop codons in error measure calculations as opposed to the fixation or fluctuation of the stop codon positions, themselves. This conclusion garnered further support from the fact that there exists great differences in ranking between modes 2 and 3, where t_{25}=3.95 and p=0.0005, with the SGC ranking lower when stop codons were included compared to when stop codons were excluded.

This led Degagne (2015) to hypothesize that the error minimizing capacity of the SGC is greater when stop codons are included in simulations, since many HGCs will possess greater amounts of stop codons than the SGC. To test this hypothesis, Degagne

(2015) conducted an analysis using 1000 HGCs and polar requirement as the property analyzed. As was anticipated, SGC ranking and the number of stop codon mutations were not significantly correlated when stop codons were excluded from error measure calculations. A strong, positive correlation was observed, however, when stop codons were included in error measure calculations (Degagne 2015: Figures 8, 9). It was proposed that the small size of the stop codon blocks compared to most amino acid blocks, in addition to the 0 value of the stop codons imparting large effects due to property value distribution of polar requirement amid the amino acids, explains the low ranking of the SGC in the error measure calculations.

Finally, Degagne (2015) utilized AGCT to test whether optimization of the SGC was best explained by coevolution theory or by four-column theory while taking into consideration the constraints that are imposed by both theories. While the coevolution theory states that new amino acid entries were constrained to codons that belonged to their biosynthetic precursor (Wong 1975, 2005), four-column theory proposes that new amino acid entries were constrained to codons that had been assigned previously to an amino acid with the most similar physico chemical properties as the new entrant (Higgs 2009). As noted in Degagne (2015), confining the shuffling of amino acid identities between codon block positions that are thought to have shared a parent block in a PGC is one method to analyze optimization in the SGC. In the case of coevolution theory, this meant that amino acid identity rearrangement was confined to codon blocks that are hypothesized to have been part of the same biosynthetic group (as was performed in Gilis et al. 2001 and Goodarzi et al 2004). In the case of four-column theory, this meant that amino acid identity rearrangement was confined to codon blocks that share the same

column, in other words, restricted to having the same nucleotide in the second codon position. It was proposed that if the SGC achieved a moderate ranking under a particular constraint, that is to say that the observed error buffering in the genetic code was attained under conditions prescribed by a theory, then the likelihood that the theory under examination successfully explained such buffering would be appreciable.

Degagne (2015) used square distance for the distance metric for 10000 HGCs, with polar requirement being the property under consideration, partly due to its eminence in past research, making for easier comparisons with other studies (Alff-Steinberg 1969, Haig and Hurst 1991, Goldman 1993, Ardell 1998, Higgs 2009 etc.). Three tests were performed, with stop codons being included in identity shuffling and error measure calculations. The first test acted as a control, wherein amino acids were assigned pseudorandomly to any available codon block. The second test enforced the constraints of coevolution theory, with amino acids restricted to codon assignments that belonged to biosynthetic precursors. The final test imposed the constraints of four-column theory, where amino acid identities could occupy only codon blocks within the column from which they had originated. As illustrated in Degagne (2015:Figure 10), the first test (no constraints) resulted in the SGC ranking 6^{th}, the second test (coevolution theory constraints) resulted in the SGC ranking 315^{th}, and the third test (four-column theory constraints) resulted in the SGC ranking 1877^{th}. Overall, the four-column theory proved more successful in terms of partial annulling of error minimization in the SGC.

A second analysis was run to test the two theories, this time assigning weighting to transition and transversion mutations. Degagne (2015) used a 2:1 ratio for transitions to transversions, that is to say that transition mutations were given twice the weighting of

a transversion mutation, since transitions are less energetically costly to undergo. The 2:1 ratio was selected as a control, as Freeland and Hurst (1998) detected that the optimal weighting ratio for transition to transversion mutations that most favored the SGC was 3:1 and, also, no favoring of transition mutations over transversion mutations would have occurred during the origin of the PGC. Degagne (2015) argued that a weighting that was neither most nor least favorable to the SGC would offer the greatest objectivity when testing coevolution theory and four-column theory. Results indicated that, once more, the four-column theory had the greatest impact on error minimization in the genetic code with a transition to transversion bias of 2:1. The first test (no constraints) resulted in the SGC ranking 1st, the second test (coevolution theory constraints) resulted in the SGC ranking 4th, and the third test (four-column theory constraints) resulted in the SGC ranking 722nd. While four-column theory was most adept at explaining error minimization, Degagne (2015) noted that the results obtained for SGC rank when accounting for coevolution theory restrictions also were significant (μ=11.18, S.D.=2.01, Z=2.46, p=0.0069). These results may have been limited in scope due to the fact that restricting amino acid identity shuffling to blocks that are biosynthetically related is, in and of itself, insufficient to account for other types of restrictions imposed by coevolution theory.

CHAPTER 2

2.1 SGC Optimality when Considering Stop Codons

One previous computational study (Goodarzi et al. 2004) analyzed SGC optimization with mutations to stop codons included in error measure calculations,

something that previously had not been performed. By accounting for amino acid frequencies and nonsense mutation frequencies, a cost function was developed, based on the function presented by Gilis et al. (2001) but incorporating relative amino acid frequencies: $\varphi^{HH} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} f(c) \sum_{c'=1}^{64} p(c'|c) g(a(c), a(c'))$, where $f(c) = \begin{cases} \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx, & y \le 0 \\ 1 - \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx & y > 0 \end{cases}$. Mutations to amino acids that have many codons assigned to them were given a greater weighting than were amino acids that have few codons assigned to them. Goodarzi et al. (2004) also utilized the same values for the probability of mistranslation at each codon position, defined by the term $p(c'|c)$ (Freeland and Hurst 1998, and Gilis et al. 2001). The values were as follows: 1) $p(c'|c) = 1/N$ if c' and c possess dissimilar nucleotides at the third position; 2) $p(c'|c) = 1/N$ if c' and c possess dissimilar nucleotides at the first position due to a transition mutation; 3) $p(c'|c) = 0.5/N$ if c' and possess dissimilar nucleotides at the first position due to a transversion mutation; 4) $p(c'|c) = 0.5/N$ if c' and c possess dissimilar nucleotides at the second position due to a transition mutation; 5) $p(c'|c) = 0.1/N$ if c' and c possess dissimilar nucleotides at the second position due to a transversion mutation; and 6) $p(c'|c) = 0$ for all other scenarios. Goodarzi et al. (2004) also applied the same rules for generating hypothetical genetic codes (HGCs) that were used by Freeland and Hurst (1998). The rules were as follows: 1) 21 non-overlapping codon sets defined the divisions in the SGC for the codon space, itself, with each codon set assigned to an amino acid and one set assigned to stop codons; 2) HGCs were generated through the pseudorandom assignment of amino acids to codon sets, with the stop codon set remaining unchanged in position for all HGCs.

Despite stop codons being excluded from codon block identity shuffling, Goodarzi et al. (2004) included them in error calculations. This was accomplished because triplets that mutated easily into stop codons (i.e., triplets that mutated to a stop codon through a single-nucleotide change) would have occurred less frequently in genetic material such as RNA, so HGC fitness would increase overall, as fewer triplets would have mutated easily into stop codons via mutation. A variable, K, was defined that allowed the cost function to be altered so that stop codons could be included in calculations. For example, if the variable K became –K, then the term g(a(c), a(c′)) became g(x, stop codon) due to a mutation. As K increased positively, the cost to HGCs decreased; and, as K increased negatively, the cost to HGCs increased. In this way, the frequency of nonsense mutations could be accounted for in calculations.

Goodarzi et al. (2004) also considered coevolutionary theory, by grouping 12 biosynthetically related amino acids together in pairs (Goodarzi et al. 2004:Table 2). If an amino acid pair (i.e., the amino acid encoded before mutation and the resulting amino acid encoded after mutation) were one of the 12 biosynthetically related pairs, then the cost measure would favour those two amino acids being placed near each other in a HGC. Goodarzi et al. (2004) noted that this analysis pertained to a period of time in which all twenty amino acids already would have been assigned to the SGC. Finally, a Z-value was defined by Goodarzi et al. (2004), as $Z = \frac{\varphi_{sgc} - \mu}{\sigma}$; no HGC was found to rank lower than the did SGC, yet distributions of error measures for HGCs closely followed Gaussian distributions (Goodarzi et al. 2004: Figures 2 and 5). In this case, μ represented the average rank of HGCs, σ represented the standard deviation of those ranks, and φ_{sgc}

represented the rank of the SGC. As can be deduced, higher values of Z lead to decreased odds of a randomly generated code scoring higher than the SGC by chance alone.

Goodarzi et al. (2004) applied their fitness function, φ^{HH}, using a $PAM_{74\text{-}100}$ scoring matrix as a cost function for 2×10^9 pseudorandomly generated HGCs, and found that none ranked higher than did the SGC when stop codons were included in calculations. This stands in minor contrast to results obtained by Gilis et al. (2001), where two out of every 10^9 HGCs scored higher than did the SGC (Goodarzi et al. 2004:Figure 2). In a second set of experiments, Z-values were calculated for two simulations that used a $PAM_{74\text{-}100}$ and Mutation Matrix in the cost function and plotted against varying K-values (Goodarzi et al. 2004:Figure 3). For each of the scenarios, 4×10^7 HGCs were generated prior to Z-values being calculated. The greatest Z values occurred when K = 4.5 with the $PAM_{74\text{-}100}$ scoring matrix as the cost function and when K = 3.0 for mutation matrix, implying that the HGC had a lower relative cost and a lower probability of scoring higher than did the SGC by chance alone. Therefore, the average cost of a codon being mistranslated into a stop codon (i.e., a nonsense mutation) was -4.5 in $PAM_{74\text{-}100}$ and -3.0 in mutation matrix.

Goodarzi et al. (2004) conducted a third set of tests in which Z-values were plotted against K-values, which remained fixed, and the Inc term (included in $g(a, a') = Inc \mid g(a, a') + g(a, a')$) for cost measure between biosynthetically related amino acids, which was assigned different values. When using a $PAM_{74\text{-}100}$ scoring matrix, the SGC displayed no optimization whatsoever, with Z-values decreasing as Inc approached 0.1 with K = 0. This implied that greater probabilities of HGCs ranking higher than did the SGC due simply to chance alone. The same trend was observed when using K = 4.5 and a

$PAM_{74\text{-}100}$ scoring matrix, with Z-values decreasing as Inc approached 0.1. Using the mutation matrix with K = 3.0 resulted in Z-values increasing as Inc increased, so the mutation matrix method appeared to favour those HGCs that had low incidence of nonsense mistranslations among biosynthetically related amino acid pairs.

Lastly, Goodarzi et al. (2004) generated another 1.2×10^9 HGCs and the fitness function, φ^{HH} and nonsense mistranslations were applied to calculations in two different distributions. The first distribution utilized the PAM74-100 scoring matrix and K = 4.5 for nonsense mistranslations, whereas the second distribution utilized the mutation matrix and K = 3.0 for nonsense mistranslations. No HGCs were observed to rank higher than did the SGC in either score distribution (Goodarzi et al. 2004:Figure 5).

2.2 Non-Bonded Interactions

Degagne (2015) observed that long-range non-bonded energy (El) was the most conserved property in the SGC when stop codons were included in error measure calculations. Three explanations were offered for this result: 1) the SGC became buffered against errors correlated to El; 2) the SGC became buffered against errors correlated to polar requirement, with El buffering being correlated strongly with polar requirement; and 3) the SGC became buffered against errors involving both polar requirement and El, as both properties may be principal guiding forces in the folding of proteins. Although there exists some discrepancy as to whether van der Waals should be considered long-range or short-range forces (IUPAC 1997, Tao and Perdew 2005, Ortmann et al. 2005, Schwabe and Grimme 2007, Kuhn and Rahe 2014), non-bonded forces contribute to the total non-bonded energy in a molecule. Given the high optimization of the SGC for the

property El, it seems prudent to review van der Waals forces and the degree to which they contribute to amino acid interactions and protein folding.

van der Waals forces can be divided generally into three different types of interaction. The first interaction type is the Keesom interaction, otherwise known as the interaction between two permanent dipoles. This type of interaction is governed by electrostatic attraction or repulsion. For example, HCl has a positive and negative pole, and will attract or repel other HCl molecules depending on their orientation relative to one another. This is considered the weakest of the van der Waals forces (Leite et al. 2012).

The second interaction type is the Debye interaction, otherwise known as induction or the interaction between a permanent dipole and an induced dipole. A permanent rotating dipole will cause an induced dipole in another element or molecule, resulting in either an attractive or repulsive force (Leite et al. 2012). For example a compound with a permanent dipole like HCl, will repel from the Cl-side or attract from the H-side, the electron cloud of Xe and cause an induced dipole in the Xe atom.

The third type of interaction is the London dispersion interaction, also known as the interaction between induced dipoles. This type of interaction is governed by the total instantaneously occurring dipole moments across all atoms and molecules in a system. For example, the electron cloud densities of atoms may undergo random fluctuations due to the various competing attractive and repulsive forces between elements and compounds. In general, London dispersion forces tend to increase with increasing surface area contact between molecules (Leite et al. 2012).

2.2.1 Density Functional Theory

Accurately measuring the precise contribution of van der Waals forces has remained a difficult endeavour. Research conducted by Grimme (2004) utilized Density Functional Theory (DFT-D) to describe empirically the large-scale contribution of van der Waals forces that arise from small, interacting complexes, such as Ne, H_2O, CH_4, NH_3, CH_3F, N_2, F_2, formic acid, ethane, ethine, benzene, naphthalene, pyrene, coronene dimers, and four stacked & H-bonded nucleotides (A-T and C-G). Grimme (2004) noted that electrostatic and exchange-repulsion forces (Keesom and Debye forces) are easy to describe using mean-field theory (the study of complex systems through the analysis of simplified models), but including many complexes in a system, such as in coupled cluster techniques, results in incredibly long time periods for computational calculation. The time required to perform these computational calculations becomes extensive even for moderate numbers of interacting bodies in a system, making studying many-body systems relatively inefficient. While there exist inexpensive alternatives, such as Moller-Plesset perturbation theory (MP2), they often have the drawback of overestimating binding energy and underestimating equilibrium distances between interacting molecules. BLYP and PBE were the two gradient-corrected density functions used to compute the dispersion present in different complexes in Grimme (2004).

While the DFT is the most frequently applied tool in studying the structure of molecules, recent knowledge has demonstrated that gradient corrections to density functions are still unable to depict accurately the effect of dispersion forces. However, several publications have demonstrated that a method to resolve this would be the addition of an empirical term, in the form C^6R^{-6}, to the existing energy terms described

by DFT, where C_6 = dispersion coefficients and R = distance between atoms. The total dispersion energy of a given system is defined as $E_{MF\text{-}D} = E_{MF} + E_{disp}$, where E_{MF} is the mean-field energy (obtained with DFT) and E_{disp} is the corrected empirical dispersion. The E_{disp} is further defined as $E_{disp} = -s_6 \sum_{i=1}^{N_{at}-1} \sum_{j=i+1}^{N_{at}} \frac{C_6^{ij}}{R_{ij}^6} f_{disp}(R_{ij})$, where N_{at} is the total number of atoms being considered in the system, C^{ij}_6 is the dispersion coefficient for a pair of atoms *i* and *j*, s6 is a factor for global scaling, and R_{ij} is the distance between atoms. Grimme (2004) also included a damping function, $f_{disp}(R) = \frac{1}{1+e^{-d(R/R_0-1)}}$, due to the fact that near-singularities occur at smaller values of R. In the damped function, R_0 represents the sum of van der Waals radii from atoms. Finally, the dispersion coefficient used is merely a set of averages, $C_6^{ij} = 2\frac{C_6^i C_6^j}{C_6^i + C_6^j}$. Using these methods, Grimme (2004) empirically analyzed the van der Waals interactions among many systems, including the four nucleotides found in DNA.

Grimme (2004) selected stacked (S) and H-bonded (Watson-Crick, WC) structures for analysis (Grimme 2004:Figure 4, 26-29). The values obtained for atomic distances and change in dispersion energy were compared to reference energies from a past study (Jurecka and Hobza 2003). At the DFT-D-BLYP level, the observed values for the nucleotide pairs did not differ significantly from a reference data set, with the greatest energy difference (0.4 kcal/mol) being found in stacked G-C pairs (Grimme 2004:Table 3). Grimme (2004) additionally noted that the correction term for dispersion resulted in improved binding energies when considering H-bonded nucleotide pairs by 4-5 kcal/mol. This is taken as evidence that a considerable amount of dispersion energy is present in the

system, even when accounting for H-bonding. At the DFT-D-PBE level, stacked conformations of nucleotide pairs were somewhat underbound when compared to H-bonded conformations. For stacked G-C pairs in DFT-D-PBE, no minimum optimal interaction energy was observed; instead, the molecular geometry appeared to prefer an H-bond governed structure. Grimme (2004) noted that this also points to PBE functions tending to disfavor stacked interactions between nucleotide pairs, but stacked G-C pairs overall, in BLYP and PBE functions, tend to prefer H-bonding with minimal dispersion energy. This is observed particularly in the nucleotide structures for DFT-D-BLYP and MP2 methods, with extensive tilts in the molecular planes of guanine and cytosine relative to each other. The close proximity of the NH_2 and C=O groups in G-C pairs in both stacked conformations and H-bonded conformations is likely the basis for observed molecular tilting. Grimme (2004) claimed that the DFT-D model was accurate to within 5-10 pm for intermolecular systems where van der Waals forces contribute significantly to stability. Overall, the study demonstrates that correcting the DFT functions, BLYP and PBE, by accounting for dispersion force effects, enabled accurate prediction about the binding energy and structural conformation in various molecular complexes that are weakly bonded.

2.2.2 van der Waals and Secondary Protein Structures

Further research (Street and Mayo 1999) found that the inclination of the natural amino acids to form secondary structures, particularly β-pleated sheets, is governed by van der Waals interactions. A van der Waals energy function was used to model dipeptides, and amino acids were plotted on Ramachandran plots (graphic representations of the available energies of amino acid residues in dihedral angles for protein backbones).

Helmholtz free energy and entropy, which resulted from placing amino acids in what was normally a β-pleated sheet region, was then calculated. Each of the 20 natural amino acids were defined as X_{aa} and modeled in dipeptide structures flanked by alanine; taking on the conformation Ala- X_{aa} -Ala. Alanine was chosen as the flanking amino acid due to its high predisposition to forming α-helices as a result of the lack of entropy loss in its side chains as the backbone is constricted into a helix shape (Hermans, Anderson, and Yun 1992). Any side chain wherein van der Waals energy for an atom exceeded a threshold value of 2.5 kcal/mol was rejected on the basis that it underwent self-collision.

The correlation between the expected propensity for an amino acid to develop β-pleated sheets and its observed (average normalized) propensity for developing β-pleated sheets was plotted, with comparisons being drawn between the change in entropy of β-sheet regions and the change in Helmholtz energy of folding β-pleated sheet regions (Street and Mayo 1999:Figure 1). Overall, the amino acids that contained two non-hydrogen groups on their C-beta carbon atom (Ile, Val, and Thr) displayed the greatest tendency for forming β-pleated sheets, followed closely by the aromatic amino acids (Phe and Tyr). Ala and Asp showed the lowest tendency for developing β-pleated sheets, as expected due to the increased preference of Ala forming α-helices. The authors noted that Gly and Pro were not included in the study as the experimental proclivity of these amino acids for forming β-pleated sheets was unclear. Asn was the only amino acid that lay significantly away from the best-fit line, which the authors noted as unusual due to the high degree of structural similarity between Asn and Asp. Normally, this would indicate that Asn and Asp should have similar van der Waals energies, thus, similar β-pleated sheet formation tendencies. Street and Mayo (1999) proposed that the slightly greater

tendency of Asn to form β-pleated sheets compared to the other amino acids may result from hydrogen bonding.

Overall, the study quantitatively demonstrated that the primary forces behind the development of β-pleated sheets in a dipeptide environment are the van der Waals forces, specifically the prevention of steric interactions between the side-chain of an amino acid and the local backbone of the amino acid. The inclusion of Coulomb forces and effects from neighbouring solvent molecules did not significantly impact propensity for β-pleated sheet formation. The results presented in that study may be interpreted as evidence for the increasingly important role of van der Waals forces between interacting amino acids, which also is significant when considering the high optimization of long-range non-bonded energy observed in the SGC.

2.3 Modifying AGCT and Further Methods

The previously developed computational algorithm, AGCT, from Degagne (2015) was modified to suit the research efforts of this study. All modifications and developments to the AGCT algorithm were done using the computing software *Mathematica* (Wolfram Research, Inc. 2010; v. 7.0.1.0, v. 11, cloud v, 1.34.1). The basic premise of the program remains intact, with HGCs being generated via shuffling of triplet identities among codon blocks and computing changes for all possible point mutations to codons by treating them as distances. A total of nine possible point mutations can be performed on each codon, with mutations to stop codons having the option of being included or excluded. For the purposes of this study, mutations to stop codons were included in error measure calculations. A single, average distance value is determined for

each HGC by computing all distances between codons and calculating a mean value. When an HGC is returned with lower rank (e.g., lower values), it possesses increased buffering against errors in transcription and translation errors. All distances are sorted by AGCT to locate the distance for the SGC among the distribution and provide a ranking for it in relation to the HGCs.

A total of 54 quantitative amino acid properties (e.g., hydrophobicity, polarity, long-range non-bonded energy) were assigned appropriate values from existing literature (Degagne 2015). This allowed comparisons between the property value distance of two amino acids in the SGC that were associated with codons with property value distance of some other two amino acids in the HGC that were associated with those same codons. Amino acids assigned to codon blocks in an HGC were said to be more dissimilar than were amino acids assigned to the same codons in the SGC if the difference between the property values increased. However, a problem is introduced by symmetry when dealing with all possible pairwise comparisons. AGCT enables users to evade this issue by allowing user to calculate the absolute difference or the square difference between property values.

One of the most notable features in the current version of AGCT was the ability to assign three degrees of penalties to nonsense mutations. AGCT previously treated stop codons in the exact same manner as it did amino acids, with each stop codon having a default value of 0 for all 54 properties. This can cause significant impact depending on the specific property being analyzed and overall structure of an HGC. The degrees of penalties that were assigned to nonsense mutations were achieved by calculating the absolute pairwise differences for any property being analyzed, and organized into

minimum, zero, and maximum groups. Minimum penalties were obtained by taking the minimum absolute pairwise differences of the property being considered between the stop codons and the other amino acids. Zero penalties were used as a control, wherein any change between an amino acid and a stop codon was set at zero. Maximum penalties were obtained by taking the maximum absolute pairwise differences of the property being considered between the stop codons and the other amino acids. This allowed minimum and maximum distances for any property to be determined and allotted for error measure calculations involving stop codons. Mutations to stop codons would be more costly when maximum penalties were assigned and less costly when zero or minimum penalties were assigned.

The updated AGCT also allows for different weightings of transitions and transversions. For this inquiry, a weighted ratio of 3:1 was used to weight transitions. This ratio was chosen despite the fact that each individual nucleotide in a codon has a 2/3 chance of undergoing an unbiased mutation that would result in a transversion. In reality, transitions are much less energetically costly and occur at greater frequencies, as replacing a purine with another purine (double-ring to double-ring) or a pyrimidine with another pyrimidine (single-ring to single-ring) is energetically more favourable than is an interchange between purine and pyrimidine (Morton 1995, Ebersberger et al. 2002). This also means that transition mutations are less likely to result in non-synonymous substitutions. Therefore, a 3:1 ratio favouring transitions was used because of its maximal efficiency for the standard genetic code (as noted by Freeland and Hurst 1998) and due to the realistic prevalence of transitions over transversions.

Two further options are available to users when generating HGCs with AGCT, one of which was of paramount importance to this study. When specified, the first option will return HGCs in which the amino acid identities have been shuffled but the stop codon identities remain fixed in their respective blocks. The second option, utilised in this study, returned HGCs wherein stop codon identities were shuffled along with the amino acid identities. Notably, AGCT was the first computational tool that allowed stop codon identities to be shuffled along with amino acid identities, and the second to allow for stop codons to be included in error measure calculations.

One constraint that remains conspicuous is the fact that the two vertically contiguous stop codon blocks (i.e., UAA and UAG) were treated as a singular identity, separate from the stop codon block in the fourth column (UGA) of the code, similar to how Ser and Arg were assigned to non-contiguous codon blocks but were still treated as singular entities. The codon blocks for Ser and Arg, as well as the stop codon blocks in the third column and fourth column, were treated as singulars in this study for various reasons. Ser and Arg both contain a primary block of four codons and a secondary block of two codons, totalling six codons for each amino acid. The secondary set of two codons for Ser and secondary set of two codons for Arg occupy the same block in the fourth column of the SGC. In this fourth column block of secondary codons, those that code for Ser (i.e., AGU and AGC) separate the primary Arg block from the secondary Arg block. In treating the Ser and Arg blocks as singulars, the number of non-overlapping codon sets for amino acids remains 20. Another reason the Ser and Arg blocks were treated as singulars was to allow for easier comparisons to past results in the literature. The vast majority of research involving the generation of HGCs kept the Ser and Arg codon blocks

as singulars; following suit in this study makes drawing relationships and contrasting between results a simpler task.

The stop codon blocks were treated in a similar manner, with those in the third column (i.e., UAA and UAG) being treated as a singular and distinct from the stop codon block in the fourth column (UGA). The degree of contiguity between the stop codons of the third column block is greater than that of the fourth column block. For example, codon UAA is able to undergo a single point mutation at the third position to become UAG. Not only are third position mutations the most likely to occur (Woese 1965), but also the transition from A to G is a purine-to-purine mutation and energetically favoured compared to a transversion. Therefore, the stop codons of the third column can be said to have the highest possible degree of contiguity with each other.

For UAA to undergo a point mutation that results in UGA there must be a transition at the second position. While transitions are energetically inexpensive, mutations are least likely to occur at the second position due to the fact that degeneracy is least determined at this position. Similarly, the codon UAG requires both a second and third position transition. Additionally, several organisms have been documented as having amino acids encoded by codons that normally serve as stop codons; specifically, selenocysteine is encoded by UGA (Low and Berry 1996) while pyrrolysine is encoded by UAG (Krzycki 2005). Considering the increased redundancy possessed by the codon UAG due to its contiguity with UAA and the fact that selenocysteine is encoded by only a single codon, UGA, considering UAA and UAG as a singular block and UGA as a separate block seemed appropriate. In this way, 22 non-overlapping codon sets were used for generating HGCs, with 20 sets for amino acids and 2 sets for stop codons.

Due to the enormous possible amount of HGCs that can be generated when including stop codons in identity shuffling ($2.59*10^{22}$), a smaller representative sample size was used for time-efficient analysis. AGCT was used to generate 10 sets of 2500 HGCs for each of the 54 properties individually. Stop codons were included in amino acid identity shuffling as well as error measure calculations. In total, each property was tested three times using 10 sets of 2500 HGCs with minimum, zero, and maximum stop codons penalties being assigned, respectively. The SGC was ranked against the HGCs for each set in order to determine the level of optimization for a particular property when accounting for stop codons, stop codon penalties, and transition-transversion bias. AGCT was updated to allow for comparisons between HGCs and the 1-in-a-million code described by Freeland and Hurst (1998). The error measure used in all calculations was the squared difference (SD), which takes the square of the difference between two values between amino acids for a given property. SGC ranking can be influenced by larger differences due to squaring, but ranking is also dependent on HGC structure. Similar amino acids are organized in columns in the SGC, so most errors in transcription or translation at the first and third position of codons yield identical or similar amino acids. If an HGC contained more significant differences in property values among amino acids that share a column, then it would possess greater error measures than those obtained for the SGC.

Finally, the average ranking was calculated for the SGC, both for individual sets and among all properties. For individual sets, an average SGC ranking was calculated for minimum stop codon penalties, zero stop codon penalties, and maximum stop codon penalties among individual properties. Additionally, a table was constructed displaying

all 54 properties, from most to least optimized, in order to determine which property was the most buffered against error when all aforementioned factors were taken into account. Lastly, an additional table was constructed displaying the hypothesized order of amino acid additions to the genetic code predicted by coevolution theory, four-column theory, and first principles theory.

2.4 Results

Overall, the SGC achieved lower ranks when applying maximum penalties to stop codons and higher ranks when applying minimum or zero penalties. Notable exceptions to this occurred when testing certain properties, such as property 1 (hydrophobicity), where the average rank for maximum penalties was 36.4 and average rank for minimum penalties was 48.5. In total, 17 properties had instances of the SGC in maximum penalty simulations ranking lower, on average, than in minimum or zero penalty simulations. Fourteen of the properties where maximum penalties scored lower had differences between average rank <50 (properties 1, 7, 8, 14, 15, 34, 38, 39, 40, 41, 45, 47, 48, 51, and 54). Two of the properties where maximum penalties scored lower had differences between average rank >50 (properties 49 and 52). Several properties displayed significant differences depending on the applied penalty. For example, property three (absolute entropy) scored an average rank of 78.7 and 62.1 for minimum and zero penalties, respectively, but 975.5 for maximum penalties.

The most drastic examples of minimum and zero penalty simulations ranking lower than maximum penalty simulations occurred in properties 10 and 16. For property 10, average ranks when applying minimum and zero penalties were 293.3 and 300.9,

respectively, while average rank when applying maximum penalty was 1899.4. For property 16, average ranks when applying minimum and zero penalties were 53 and 50.6, respectively, while average rank when applying maximum penalty was 2011.7. As was first observed in Degagne (2015), long-range non-bonded energy was the most conserved property, with an average ranking of 1.2 when applying maximum penalty and 1.1 when applying both zero and minimum penalties. This was closely followed by the previously observed most conserved property, polar requirement, which scored an average ranking of 1.8 when applying maximum penalty, 1.4 when applying zero penalty, and 1.1 when applying minimum penalty. The most conserved properties (i.e., those that returned non-averaged ranks <5 for all three penalty options) were polar requirement, chromatographic index, long-range non-bonded energy, and unfolding enthalpy change of hydration.

Figures 1, 2, and 3 display the average rankings of the SGC against 2499 HGCs for maximum, zero, and minimum penalties, respectively, over the 54 properties, with standard deviation error bars. Long-range non-bonded energy was observed to be the most conserved property overall, with an average maximum penalty rank of 1.2, average zero penalty rank of 1.1, and average minimum penalty rank of 1.1. Polar requirement, often cited as the most conserved property, was the second-most conserved property overall, with an average maximum penalty rank of 1.8, average zero penalty rank of 1.4, and average minimum penalty rank of 1.1. Table 1 displays theorized chronologies of amino acid appearance in the genetic code according to coevolution theory, four-column theory and first principles theory, with first-principles theory outlining specifically when stop codons would have been expected to emerge. Table 2 displays properties 1 through 54 and their corresponding symbols.

2.5 Discussion

Long-range non-bonded energy was observed to be the most conserved property, in agreement with the results found in Degagne (2015). Studies on non-bonded interactions have outlined the great importance that non-bonded energy in protein folding and secondary structure formation between amino acids, but to claim the SGC became organized around long-range non-bonding energy of amino acids would be a gross over simplification. Non-bonded interactions within molecules are vast, fluctuating networks of attractive and repulsive forces that inherently are related to many other properties (such as hydrophobicity, surface area in folded vs. unfolded conformations, atomic radius and electron orbital shape). For this reason, long-range non-bonded energy may be so highly conserved because of its close relationships to several significant properties, resulting in a cumulative effect that leads to extremely high conservation.

It is noteworthy that polar requirement possessed an equivalent average rank value for the minimum stop codon penalty option and a lower average rank value for maximum penalty options, implying that maximum stop codon penalties have greater overall impact on polar requirement of amino acids, perhaps due to property value distribution of polar requirement among the amino acids. The average ranking for unfolding enthalpy change of hydration was third for maximum and minimum penalties but fourth for zero penalties. This implied that stop codon penalties were not particularly detrimental to this property and the SGC is highly buffered for it. Given the importance of non-truncated proteins and properly folded proteins to the overall structure and function of a genetic code and its ability to replicate, it is of paramount importance to be buffered against effects of nonsense mutations. Such mutations will result in truncated or

malformed proteins and will vastly effect the change in enthalpy of folded and unfolded states. For example, a nonsense mutation can result in a shortened protein that improperly folds, causing non-polar side chains to be exposed and reducing the interaction capabilities and general stability of the molecule. Therefore, it seems reasonable that properties such as unfolding enthalpy change of hydration are strongly buffered when considering stop codon effects.

Curiously, there were instances of average maximum penalty ranks scoring lower values than average zero penalty ranks or average minimum penalty ranks. Specifically, this occurred in properties 1 (hydrophobicity), 7 (surrounding hydrophobicity), 8 (polarity), 14 (refractive index), 15 (normalized consensus hydrophobicity), 34 (combined surrounding hydrophobicity), 38 (Gibbs free energy change of hydration for unfolding), 39 (Gibbs free energy change for denatured protein), 40 (Gibbs free energy change for native protein), 41 (unfolding enthalpy change of hydration), 45 (unfolding entropy change of hydration), 47 (Gibbs free energy change), 48 (unfolding enthalpy), 49 (unfolding entropy changes of the chain), 51 (hydrophobicity scale), 52 (hydrophobicity scale B), and 54 (fraction of surface area lost when folded). While average maximum penalties tended to score better than minimum or zero penalties for these properties, many of these were instances wherein outliers impacted the average and caused the maximum penalty rankings to be slightly lower than their minimum or zero counterparts. For example, in properties 1, 7, 8 , 14, 15, 34, 38, 39, 40, 41, 45, 47, 48, 51, and 54, the differences in average ranks between maximum penalty simulations and minimum or zero penalty simulations was <50, so the lower value rankings of maximum penalty simulations may be attributed to outliers in the data when calculating the average rank.

However, properties 49 and 52 had differences between average maximum rank and minimum or zero rank that were >50. Property 49 possessed an average rank of 272.1 for maximum penalty simulations but an average rank of 343.9 and 329.5 for minimum and zero penalty simulations, respectively. Similarly, property 52 possessed an average rank of 483.8 for maximum penalty simulations but an average rank of 566.8 and 564.5 for minimum and zero penalty simulations, respectively. Again, property value distribution among the amino acids may be the culprit for this result, as one would expect that increasing the penalty for nonsense mutations should result in an increased rank value for the property. In this case, it may be a matter of the values for maximum absolute pairwise differences and minimum absolute pairwise differences being very close to each other for properties 49 and 52. This could realistically result in scenarios where maximum penalty simulations may outscore minimum or zero penalty simulations. Also, it is important to take the magnitude of the values into account. Even though maximum penalty scenarios outscored minimum and zero penalty scenarios for properties 49 and 52, they still only achieved an average rank between 340 and 567, which is far less than many of the more buffered properties in the SGC.

The greatest difference between maximum, zero, and minimum penalty simulations occurred when analyzing properties 10 (equilibrium constant with reference to the ionization property of the COOH group) and 16 (short and medium range non-bonded energy). Property 10 possessed an average rank of 1899.4 for maximum penalty simulations and an average rank of 293.3 an 300.9 for minimum and zero penalty simulations, respectively. Property 16 possessed an average rank of 2011.7 for maximum penalty simulations and an average rank of 53 and 50.6 for minimum and zero penalty

simulations, respectively. These data suggest that increasing penalties for nonsense mutations has an exacting detrimental effect on both the ionization of COOH groups in amino acids and the short range and medium range non-bonding forces of amino acids. Properties 10 and 16 may be intricately linked to each other in this regard as all amino acids are considered to be zwitterions (dipolar ions), a neutral molecule that contains both negatively charged groups and positively charged groups.

Nonsense mutations that truncate proteins can result in a molecule with insufficient numbers of COOH groups for non-bonding interactions or imbalances between positive and negative charges in the overall molecule. Additionally, the amine groups in amino acids can undergo deprotonation reactions with the carboxylic acid groups of other amino acids, resulting in intramolecular stability. By considering the increased negative consequences of maximum penalties for nonsense mutations, one can propose that intramolecular structure in proteins is compromised due to a lack of balance between numbers of amine groups and carboxylic acid groups as a consequence of protein truncation. This is further reinforced by the fact that the scores for minimum and zero penalties for properties 10 and 16 are 1 and 2 orders of magnitude lower, respectively, than the scores for maximum penalty simulations.

The analyses, and subsequent results, presented herein also have significance when considering theories on the origins of the genetic code. Table 1 depicts the theorized order in which amino acids appeared in the code according to coevolution theory, first principles theory, and four-column theory. The coevolution theory and four column theory disagree in which amino acids were first established within the genetic code, while the four column theory simply uses an average consensus of many theories to

establish an amino acid chronology. The arrival and assimilation of stop codons into the code is largely ignored by these theories, with only the first principles theory (Trifonov 2004) directly addressing their arrival within the context of a codon chronology.

According to first principles theory, the termination codons made their initial appearances at steps 15, 25, and 31 as a result of complementary copying of nucleotides. In first principles theory, the first stop codon (UGA) did not appear until after 10 amino acids were incorporated into the code and the final stop codons (UAA and UAG) did not appear until after 12 amino acids were incorporated into the code and all but two codons had been assigned to amino acids. Adding the stop codons at such a late stage can be both beneficial and detrimental. It is beneficial in the sense that the addition of stop codons can vastly impact decrease translation costs, as an entire strand of genetic it can be detrimental material need not be translated to make a specific protein from a certain gene. However, it can be detrimental in the sense that adding the stop codons so late into genetic code may fundamentally effect the structure of a genetic code, as is hinted at in our study. For example, some properties remain relatively unaffected by stop codons and their associated penalties and achieve low value ranks (i.e., combined surrounding hydrophobicity, polar requirement) while other properties are heavily influenced by stop codons and their associated penalties and achieve high value ranks (i.e., thermodynamic transfer hydrophobicity, alpha-helix tendency).

Many theories do not account for, or outright ignore, the effects that stop codons can impart onto the genetic code, especially for specific properties. Generally speaking, our analyses indicate that the lower a rank is for a certain property, the less it is affected by stop codon effects, while the higher a value rank is for a certain property, the more it

is influenced by stop codon affects. Given these results, it is becoming increasingly clear that ignoring stop codons in structural analyses of the genetic code is no longer sensible. While close approximations can be made in most cases when ignoring stop codons, including them fully in future research will allow the most consistently accurate models of the genetic code and its origins to be deduced.

The next logical step would be to use the current AGCT to test certain theories of the origin of the SGC in order to determine which theory most accurately explains error buffering in the genetic code. This would be done by taking into account the constraints imposed by both theories while simultaneously including stop codon shuffling and associated stop codon penalties. Both coevolution theory and four-column theory propose that there existed a PGC prior to the SGC, with the PGC hypothesized to possess fewer amino acids than the current SGC (Wong 1975, Wong 2005, Higgs 2009). Coevolution theory proposes that new amino acids that were added to the genetic code were restricted to codons that belonged to their biosynthetic precursors, while four-column theory proposes that new amino acids that were added to the genetic code were restricted to codons that had previously been assigned to the amino acid with the highest degree of physiochemical similarity as the newcomer.

In testing coevolution theory, identity shuffling should be confined to codon blocks that are part of the same biosynthetic family (Wong 2005:Figure 1). In testing four-column theory, identity shuffling should be confined to codon blocks that are part of the same column (Higgs 2009:Figure 2 and 3). By using the newest version of AGCT, nonsense mutations may be completed accounted for under various different parameters, for example, one could test the error buffering capacity of coevolution theory when fully

considering stop codons, one or multiple properties, different transition:transversion ratios, and two different distance metrics. Ideally, the SGC should achieve a moderate ranking under the constraints of the theory that most accurately predicts the observed error minimization in the genetic code, but if the SGC ranked lower, then the theory being tested is less likely to provide an accurate prediction of the observed error minimization.

Testing the first principles theory using the above method proves far more difficult as it introduces amino acids as well as codons simultaneously, building off the assumption that not all triplets would have been available. There would be no choice but to test first principles theory in stages. For example, the first two amino acids (Gly and Ala) and the first two codons (GGC and GCC) would compose the first stage of the genetic code and identity shuffling would be restricted to those two amino acids and codons. The next stage would utilise the amino acids and codons from the first stage but also include the next amino acids and codons that were hypothesized to appear at that point in first principles theory. Certain amino acids and codons appear simultaneously according to first principles theory, and this should be reflected when adding amino acid and codon sets sequentially to genetic codes to be tested. Obviously, this greatly increases the number of possible HGCs that must be tested, but one could surmise that the early genetic codes will always rank poorly compared to those that arrive later, especially when nonsense mutations are accounted for since early genetic codes will lack degeneracy.

Computational analyses of the genetic code are numerous and have been performed as far back as the late-1960s (Alff-Steinberger 1969). However, much of this

literature is purely computational and few efforts are made to connect the results from optimization analyses with biochemical principles and theories. Despite this, there is still information to be gleaned from optimization analyses regarding the origins of the genetic code. Thoroughly accounting for stop codons effects has aided in this endeavour by revealing certain properties to be highly conserved that previously were not considered important (i.e., El, C_{ph}, and fraction of surface area lost when folded). It is becoming clear that in order to accurately model the origins of the genetic code, certain properties must take precedence over others. New and existing theories should consider assigning weighting to properties based on the level of conservation of properties in the SGC when considering stop codons, and how this may effect an evolving PGC. Another challenge that remains is pinpointing when the stop codons would have been incorporated into the genetic code, as this would have subtle yet potent effects on the final structure of the SGC.

In conclusion, the results of this study indicate that increasing the penalty for stop codon mutations, while simultaneously allowing stop codons to be included in identity shuffling, results in decreased rankings for the SGC compared to HGCs for most properties, while decreasing penalties for stop codon mutations results in increased rankings for the SGC compared to HGCs for most properties. While 16 properties do score better under maximum penalties, this can mainly be attributed to outliers in the data when calculating average ranks for those properties. The analyses presented herein suggest also that non-bonded forces may play a larger role in code optimization than has been recognized previously, an observation that seems to recur when fully considering stop codon effects on the SGC. Through consideration, in greater detail, the effects of

stop codons by way of assigning maximum, minimum, and zero penalties resulted in an increasingly accurate analysis of optimization in the genetic code, allowing a more complete illustration of the origins for the code's structural organization to be ascertained. An informative follow up to the study presented here would be to further consider the effects of stop codons, including penalty options, while treating the stop codon blocks as individuals. For example, future analyses would utilise three stop codon blocks in identity shuffling, the two present in column 3 of the SGC and the one present in column 4. As a complement to this, the codon blocks for Ser and Arg may also be separated, such that the primary Ser and Arg blocks are shuffled separately from their respective secondary blocks. Alternative transition-transversion bias ratios also should be used, as the 3:1 bias present in this study most favours the SGC. Ultimately, these further analyses will help paint the clearest possible picture of the SGC structure, uncovering evidence that will aid researchers in reconstructing the complete history of the genetic code's genesis and subsequent evolution.

References

Alff-Steinberger, C. (1969). The genetic code and error transmission. *Proceedings of the National Academy of Science*, 64(2): 584-591.

Altschul, F., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17): 3389-3402.

Amirnovin, R. (1997). An analysis of the metabolic theory of the origin of the genetic code. *Journal of Molecular Evolution,* 44: 473–476.

Ardell, D.H. (1998). On error minimization in a sequential origin of the standard genetic code. *Journal of Molecular Evolution*, 47: 1–13.

Bacher, J., Ellington, A. (2001). Selection and variation of *Escherichia coli* variants capable of growth on an otherwise toxic tryptophan analogue. *Journal of Bacteriology*, 183(18): 5414-5425.

Bacher, J., Bull, J., Ellington, A. (2003). Evolution of phage with chemically ambiguous proteomes. *BMC Evolutionary Biology*, 3:24.

Bacher, J., et al. (2004). Evolving new genetic codes. *Trends in Ecology and Evolution*, 19(2): 69-75.

Barrell, B.G., et al. (1980). Different pattern of codon recognition by mammalian mitochondrial tRNAs. *Proceedings of the National Academy of Sciences*, 77(6): 3164-3166.

Bermudez, C., Daza, E., Andrade, E. (1999). Characterization and comparison of *Escherichia coli* transfer RNAs by graph theory based on secondary structure. *Journal of Theoretical Biology*, 197(2): 193-205.

Biebricher, C.K., Eigen, M. (2006). What is a quasispecies. In *Quasispecies: Concept and Implication for Virology* (1-31). Springer Berlin Heidelberg.

Bock, A. (2001). Invading the genetic code. *Science*, 292(5516): 453-454.

Budisa, N., Moroder, L., Huber, R. (1999). Structure and evolution of the genetic code viewed from the perspective of the experimentally expanded amino acid repertoire *in vivo*. *Cellular and Molecular Life Sciences*, 55(12): 1626-1635.

Cavalcanti, A., Leite, E., Neto, B., Ferreira, R. (2002). On the classes of aminoacyl-tRNA synthetases, amino acids, and the genetic code. *Origins of Life and Evolution of the Biosphere*, 34(4): 407-420.

Chambers, I., et al. (1986). The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded ‘termination’ codon, TGA. *The EMBO Journal*, 5(6): 1221-1227.

Chiusano, M., et al. (2000). Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene*, 261(1): 63-69.

Commans, S., Bock, A., (1999). Selenocysteine-inserting tRNAs: an overview. *Microbiology Reviews*, 23(3): 335-351.

Crick, F.H.C. (1963). On the genetic code. *Science*, 139(3554): 461-464.

Danchin, A. (1989). Homeotopic transformation and the origin of translation. *Progress in Biophysics and Molecular Biology*, 54: 81-86.

Davis, B.K. (1999). Evolution of the genetic code. *Progress in Biophysics and Molecular Biology*, 72(2): 157-243.

Degagne, C. (2015). A computational analysis of the structure of the genetic code (Master's Thesis).

Delarue, M. (1995). Partition of aminoacyl-tRNA synthetases in two different structural classes dating back to early metabolism: implications for the origin of the genetic code and the nature of protein sequences. *Journal of Molecular Evolution*, 41(6): 703-711.

Di Giulo, M. (1989). The extension reached by the minimization of the plarity distances during the evolution of the genetic code. *Journal of Molecular Evolution*, 29(4): 288–293.

Di Giulio, M. (1996). The b-sheets of proteins, the biosynthetic relations between amino acids, and the origin of the genetic code. *Origins of Life and Evolution of the Biosphere*, 26: 589-609.

Di Giulio, M. (1999). The coevolution theory of the origin of the genetic code. *Journal of Molecular Evolution*, 48(3): 253-254.

Di Giulio, M. (2004). The origin of the tRNA molecule: implications for the origin of protein synthesis. *Journal of Theoretical Biology*, 226(1): 89-93.

Di Giulio, M., Medugno, (1998). The historical factor: the biosynthetic relationships between amino acids and their physiochemical properties in the origin of the genetic code. *Journal of Molecular Evolution*, 46: 615-621.

Di Giulio, M. & Medugno M. (2000). The robust statistical bases of the coevoution theory of genetic code origin. *Journal of Molecular Evolution*, 50: 258–263.

Dillon, L. (1973). The origins of the genetic code. *The Botanical Review*, 39(4): 301-345.

Dillon, L.S. (1978). The genetic mechanism and the origin of life. Plenum Press, New York and London.

Dunnill, P. (1966). Triplet nucleotide-amino-acid-pairing; a stereochemical basis for the division between protein and non-protein amino-acids. *Nature*, 210(5042): 1267-1268.

Ebersberger, I., et al. (2002). Genome-wide comparison of DNA sequences between humans and chimpanzees. *The American Journal of Human Genetics*, 70(6): 1490-1497.

Edwards, M. (1996). Metabolite channelling in the origin of life. *Journal of Theoretical Biology*, 179(4): 313-322.

Eigen, M., Schuster, P. (1978). The hypercycle. A principle of natural self-organization. *Naturwissenschaften*, 65: 341-369.

Figureau, A. (1989). Optimization and the genetic code. *Origins of Life and Evolution of the Biosphere*, 19(1): 57-67.

Folk, W., Yaniv, M. (1972). Coding properties and nucleotide sequences in *E. coli* glutamine tRNAs. *Nature New Biology*, 237: 165-166.

Francklyn, C. (2003). TRNA synthetase paralogs: evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to *de novo* biosynthesis. *Proceedings of the National Academy of Sciences*, 100(17): 9650-9652.

Freeland, S.J. & Hurst, L.D. (1998a). The genetic code is one in a million. *Journal of Molecular Evolution*, 47: 238–248.

Freeland, S.J., Wu, T., Keulman, N. (2003). The case for an error minimizing standard genetic code. *Origins of Life and Evolution of the Biosphere*, 33(4): 457-477.

Giege, R., Sissler, M., Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research*, 26(22): 5017-5035.

Gilis, D., Massar, S., Cerf, N.J. & Rooman, M. (2001). Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biology*, 2(11): 49.1–49.12.

Gilvarg, C. (1962). The branching point in diaminopimelic acid synthesis. *The Journal of Biological Chemistry*, 237(2): 482-484.

Goldman, N. (1993). Further results on error minimization in the genetic code. *Journal of Molecular Evolution*, 37: 662–664.

Grimme, S. (2004). Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 25(12): 1463-1473.

Guimaraes, R.C., Moreira, C.H.C (2004). Genetic code: a self-referential and functional model. *Progress in Biological Chirality*. Paris: Elsevier. 83-118.

Guo, Q., et al. (2002). Recognition by tryptophanyl-tRNA synthetases of discriminator base on $tRNA^{trp}$ from three biological domains. *The Journal of Biological Chemistry*, 277: 14343-14349.

Haig, D. & Hurst, L.D. (1991). A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution*, 33:412–417.

Harada, K., Fox, S. (1964). Thermal synthesis of natural amino-acids from a postulated primitive terrestrial atmosphere. *Nature*, 201: 335-336.

Hendry, L.B., Bransome Jr., E.D., Hutson, M.S., Campbell, L.K. (1981). First approximation of a stereochemical rationale for the genetic code based on the topography and physiochemical properties of "cavities" constructed from models of DNA. *Proceeding of the National Academy of Science U.S.A.*, 78: 7440–7444.

Hermans, J., Anderson, A., Yun, R.H. (1992). Differential helix propensity of small apolar side chains studied by molecular dynamics simulations. *Biochemistry*, 31(24): 5646-5653.

Higgs, P.G. (2009). A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biology Direct*, 4: 16–45.

Higgs, P.G., Atwood, T. (2005). *Bioinformatics and Molecular Evolution*. Malden, MA: Blackwell Publishing.

Hinegardner, R., Engelberg, J. (1963). Rationale for a universal genetic code. *Science*, 142(3595): 1083-1085.

Ibba, M., Curnow, A., Soll, D. (1997). Aminoacyl-tRNA synthesis: divergent routes to a common goal. *Trends in Biochemical Sciences*, 22(2): 39-42.

Ibba, M., Francklyn, C. (2004). Turning tRNA upside down: when aminoacylation is not a prerequisite to protein synthesis. *Proceedings of the National Academy of Sciences*, 101(20): 7493-7494.

Ish-Horowicz, D., Clark, B. (1973). The nucleotide sequence of a serine transfer ribonucleic acid from *Escherichia coli*. *The Journal of Biological Chemistry*. 248: 6663-6673.

IUPAC, *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book") (1997).

Johnston,W., et al. (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science*, 292(5920): 1319-1325.

Judson, O.P. & Haydon, D. (1999). The genetic code: What is it good for? An analysis of the effects of selection pressures on the genetic code. *Journal of Molecular Evolution*, 49: 539–550.

Jukes, T.H. (1971). Prebiotic and biochemical evolution. Eds. Kimball, A. P. & Oro, J. (North Holland/Elsevier, Amsterdam, London, New York),122-147.

Jukes, T.H. (1973). Arginine as an evolutionary intruder into protein synthesis. *Biochemical and Biophysical Research Communications*, 53(3): 709-714.

Jurka, J., Smith, T.F. (1987). B-turn driven early evolution: the genetic code and biosynthetic pathways. *Journal of Molecular Evolution*, 25: 15-19.

Jurecka, P., Hobza, P. (2003) True stabilization energies for the optimal planar hydrogen-bonded and stacked structures of guanine, cytosine, adenine, thymine, and their 9- and 1-methyl derivatives: complete basic set calculations at the MP2 and CCSD (T) levels and comparison with experiment. *Journal of the American Chemical Society*, 125(50): 15608-15613.

Kisselev, L., Ehrenberg, M., Frolova, L. (2003). Termination of translation: interplay of mRNA, rRNAs, and release factors. *The EMBO Journal*, 22(2): 175-182.

Klipcan, L., Safro, M. (2004). Amino acid biogenesis, evolution of the genetic code, and aminoacyl-tRNA synthetases. *Journal of Theoretical Biology*, 228(3): 389-396.

Knight, R., Landweber, XX., Yarus, M. (2003). Tests of a stereochemical genetic code. Eds. Lapointe J, Brakier-Gingras L. *Translation Mechanisms*. Georgetown, Texas: Landes Bioscience. p 115 – 129.

Kobayashi, K., Tsuchiya, M., Oshima, T., Yanagawa, H. (1990). Abiotic synthesis of amino acids and imidazole by proton irradiation of simulated primitive Earth atmospheres. *Origins of Life and Evolution of the Biosphere*, 20(2): 99-109.

Kobayash, K., Kaneko, T., Saito, T., Oshima, T. (1998). Amino acid formation in gas mixtures by high energy particle irradiation. *Origin of Life and Evolution of the Biosphere*, 28(2): 155-165.

Kowal, A.K., Kohrer, C., RajBhandry, U. (2001). Twenty-first aminoacyl-tRNA synthetase-suppressor tRNA pairs for possible use in site-specific incorporation of amino acid analogues into proteins in Eukaryotes and Eubacteria. *Proceedings of the National Academy of Sciences*, 98(5): 2268-2273.

Krzycki, J.A. (2004). Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases. *Current Opinion in Chemical Biology*, 8(5): 484-491.

Kwok, Y., Wong, J.T.F. (1980). Evolutionary relationship between *Halobacterium cutirubrum* and Eukaryotes determined by use of aminoacyl-tRNA synthetases as phylogenetic probes. *Canadian Journal of Biochemistry*, 58(3): 213-218.

Lacey, J.C., Wickramasinghe N.S.M.D., Cook, G.W. (1992). Experimental studies on the origin of the genetic code and the process of protein synthesis: a review update. *Origins of Life and Evolution of the Biosphere*, 22(5): 243-275.

Leite, F., et al. (2012). Theoretical Models for Surface Forces and Adhesion and Their Measurement Using Atomic Force Microscopy. *International Journal of Molecular Sciences*, 13(12): 12773.

Luo, L. (1989). The distribution of amino acids in the genetic code. *Origins of Life and Evolution of the Biosphere*, 19(6): 621-631

Magliery, T.J., et al. (2003). *In vitro* tools and *in vivo* engineering: incorporation of unnatural amino acids into proteins. Eds. Lapointe J, Brakier-Gingras L. *Translation Mechanisms*. Georgetown, Texas: Landes Bioscience., 95-114.

Manssant, J. et al. (2002). Metabolic channelling of carbamoyl phosphate, a thermolabile intermediate. *The Journal of Biological Chemistry*, 277: 18517-18522.

McClendon, J.H. (1986). The relationship between the origins of the biosynthetic paths to the amino acids and their coding. *Origins of Life and Evolution of the Biosphere*, 16(3): 269-270.

Melcher, G. (1974). Stereospecificity of the genetic code. *Journal of Molecular Evolution*, 3, 121–141.

Miller, S. (1953). A production of amino acids under possible primitive earth conditions. *Science*, 117(3046): 528-529.

Morton, B. (1995). Neighbouring base composition and transversion/transition in a comparison of rice and maize chloroplast non-coding regions. *Proceedings of the National Academy of Sciences*, 92(21): 9717-9721.

Moser, J. (2001). V-shaped structure of glutamyl-tRNA reductase, the first enzyme of being tRNA-dependent tetrapyrrole biosynthesis. *The EMBO Journal*, 20(23): 6583-6590.

Nagel, G., Doolittle, R. (1995). Phylogenetic analyses of the aminoacyl-tRNA synthetases. *Journal of Molecular Evolution*, 40(5): 487-498.

Ortmann, F., Schmidt, W. G. & Bechstedt, F. (2005). Attracted by long-range electron correlation: adenine on graphite. *Physical Review Letters* 95(18): 186101.

Palm, C., Calvin, M. (1962). Primordial organic chemistry. I. Compounds resulting from electron irradiation of $C^{14}H_4$. *Journal of the American Chemical Society*, 84(11): 2115-2121.

Pelc, S.R. (1965). Correlation between Coding-Triplets and Amino Acids. *Nature*, 207: 597–599.

Pezo, V. (2004). Artificially ambiguous genetic code confers growth yield advantage. *Proceedings of the National Academy of Sciences*, 101(23): 8593-8597.

Polycarpo, C. et al. (2004). An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. *Proceedings of the National Academy of Sciences*, 101(34): 12450-12454.

Ribas, L., Schimmel, P. (2001). Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends in Biochemical Sciences*, 26(10): 591-596.

Ronneberg, T.A., Landweber, L.F. & Freeland, S.J. (2000). Testing a biosynthetic theory of the genetic code: Fact or artifact?. *Proceedings of the National Academy of Sciences U.S.A.*, 97: 13690–13695.

Roy, H., Becker, H., Reinbolt, J., Kern, D. (2003). When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. *Proceedings of the National Academy of Sciences*, 100(17): 9837-9842.

Schwabe, T., Grimme, S. (2007) Double-hybrid density functionals with long-range dispersion corrections: higher accuracy and extended applicability. *Physical Chemistry Chemical Physics* 9(26): 3397–3406.

Seligmann, H., Amzallag, N. (2002). Chemical interactions between amino acid and RNA: multiplicity of the levels of specificity explains origin of the genetic code. *Naturwissenschaften*, 89(12): 542-551.

Siatecka, M., Rozek, M., Barciszewski, J., Mirande, M. (1998). Modular evolution of the glx-tRNA synthetase. *The FEBS Journal*, 256(1): 80-87.

Shimizu, M. Molecular basis for the genetic code. (1982). *Journal of Molecular Evolution*, 18: 297–303.

Sonneborn, T.M. (1965). "Degeneracy of the genetic code: extent, nature, and genetic implications". In H. Bryson & H. J. Vogel (Eds.), *Evolving Genes and Proteins* (pp. 377–397). New York, United States of America: Academic Press.

Srinivasan, G., James, C., Krzycki, J. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science*, 296(5572): 1459-1462.

Stevenson, D.S. (2002). Co-evolution of the genetic code and ribozyme replication. *Journal of Theoretical Biology*, 217(2): 235-253.

Strassman, M., Weinhouse, S. (1952). Lysine biosynthesis in *Torulosis utilis*. *Journal of the American Chemical Society*, 74(13): 3457-3458.

Szathmary, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends in Genetics*, 15(6): 223-229.

Taylor, F.J.R. & Coates, D. (1989). The code within the codons. *BioSystems*, 22: 177–187.

Trifonov, E.N. (1999). Glycine clock: Eubacteria first, Archaea next, Protoctista, Fungi, Planta, and Animalia last. *Gene Therapy and Molecular Biology*, 4: 313-322.

Trifonov, E.N. (2000). Consensus temporal order of amino acids and evolution of the triplet code. *Gene*, 261(1): 139–151.

Trifonov, E.N. (2004). The triplet code from first principles. *Journal of Biomolecular Structure and Dynamics*, 22(1): 1–11.

Trifonov, E.N. Theory of early molecular evolution. In *Discovering Biomolecular Mechanisms with Computational Biology*. Eds. Frank Eisenhaber. Landes Bioscience, Georgetown (2004) in press.

Tumbula, D., et al. (1999). Archaeal aminoacyl-tRNA synthesis: diversity replaces dogma. *Genetics*, 152(4): 1269-1276.

Vogel, H., Davis, B. (1952). Glutamic γ-semialdehyde and Δ^{1}-pyrroline-5-carboxylic acid, intermediates in the biosynthesis of proline. *Journal of the American Chemical Society*, 74(1): 109-112.

Wachtershauser, G. (1988). Before enzymes and templates: theory of surface metabolism. *Microbiological Reviews*, 52(4): 452-484.

Wheeler, D.L., et al. (2004). Database resources of the national center for biotechnology information. Update: *Nucleic Acids Research*, 32:D35-D40.

Woese, C.R. (1965). On the origin of the genetic code. *Proceedings of the National Academy of Sciences USA*, 54: 1546–1552.

Woese C.R., Dugre, D.H., Saxinger, W. C. & Dugre, S.A. (1966). The molecular basis for the genetic code. *Proceedings of the National Academy of Sciences*, 55(4): 966–976.Jung 1978.

Wong, J.T.F. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences USA*, 72(5): 1909–1912.

Wong, J.T.F. (1976). The evolution of a universal genetic code. *Proceedings of the National Academy of Sciences*, 73(7): 2336-2340.

Wong, J.T.F. (1980). Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proceedings of the National Academy of Sciences*, 77(2): 1083-1086.

Wong, J.T.F. (1981). Coevolution of genetic code and amino acid biosynthesis. *Trends in Biochemical Sciences*, 6: 33–36.

Wong, J.T.F. (1988). Evolution of the genetic code. *Microbiological Sciences*, 5(6): 174-181.

Wong, J.T.F. (1991). Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. *Origin of Life and Evolution of the Biosphere*, 21(3): 165-176.

Wong, J.TF. (2005). Coevolution theory at the age of thirty. *BioEssays*, 27: 416–425.

Wong J.T.F., Bronskill, P. (1979). Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *Journal of Molecular Evolution*, 13(2): 115-125.

Wong,J.T.F., Xue, H. (2002). Self-perfecting evolution of heteropolymer building blocks and sequences as the basis of life. Eds. Palyi, G., Zucchi, C., Caglioti, L. *Fundamentals of Life*. Paris: Elsevier, 473-494.

Xue, H., Giege, R., Wong, J.T.F. (1993). Identity elements of $tRNA^{trp}$. Identification and evolutionary conservation. *The Journal of Biological Chemistry*, 268: 9316-9322.

Xue et al. (2003). Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene*, 310: 59-66.

Yamada, Y., Ishikura, H. (1973). Nucleotide sequences of $tRNA^{Ser}_3$ from *Escherichia coli*. *FEBS Letters*, 29(3): 231-234.

Yamao, F., et al. (1985). UGA is read as tryptophan in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences*, 82(8): 2306-2309.

Yang, C.M. (2004). Molecular versus atomic information logic behind the genetic coding contents constrained by two evolutionary axes and the Fibonacci-Lucas sequence. *Journal of Biological Systems*, 12:21-44.

Yarus, M. (1988). A specific amino acid binding site composed of RNA. *Science*, 240(4860): 1751–1758.

Figure 1

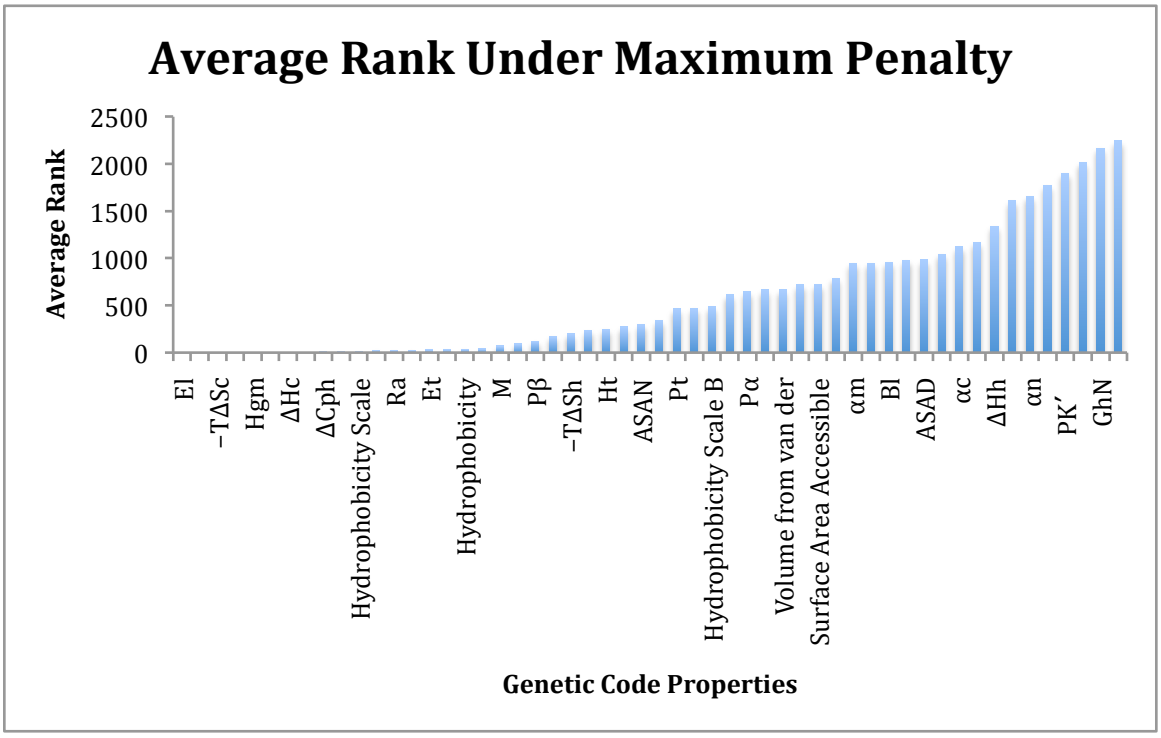


Figure 1: This bar graph represents the averaged ranks of the SGC among 10 sets of 2500 replicates for 54 different properties when maximum penalties are applied to stop codons.

Figure 2

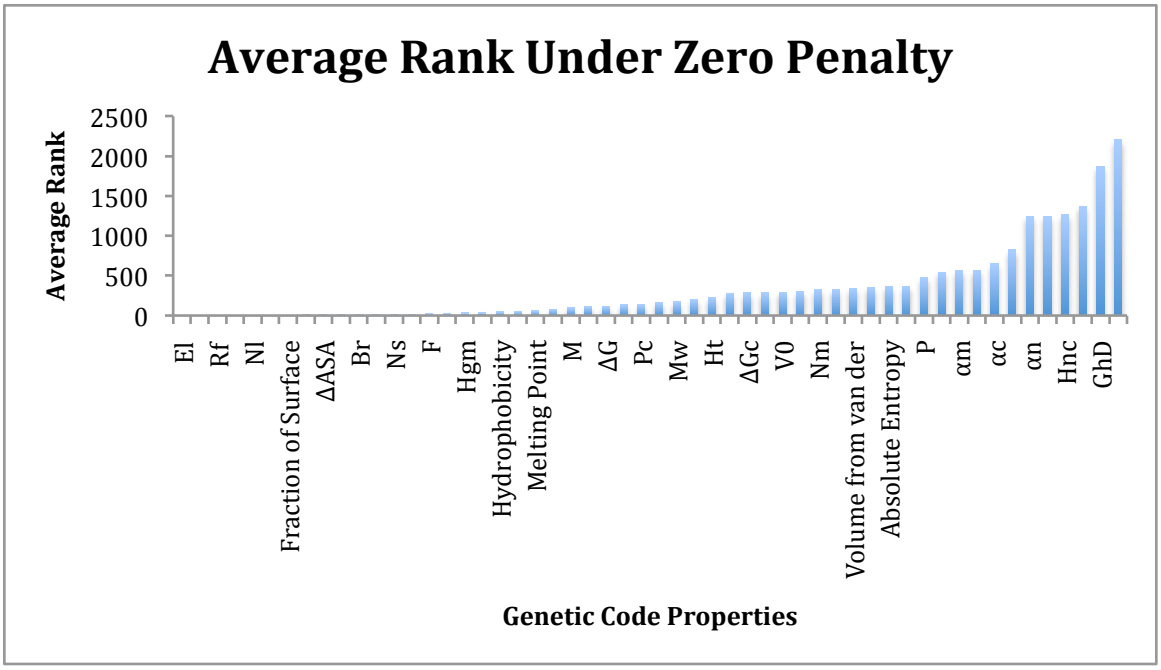


Figure 2: This bar graph represents the averaged ranks of the SGC among 10 sets of 2500 replicates for 54 different properties when zero penalties are applied to stop codons.

Figure 3

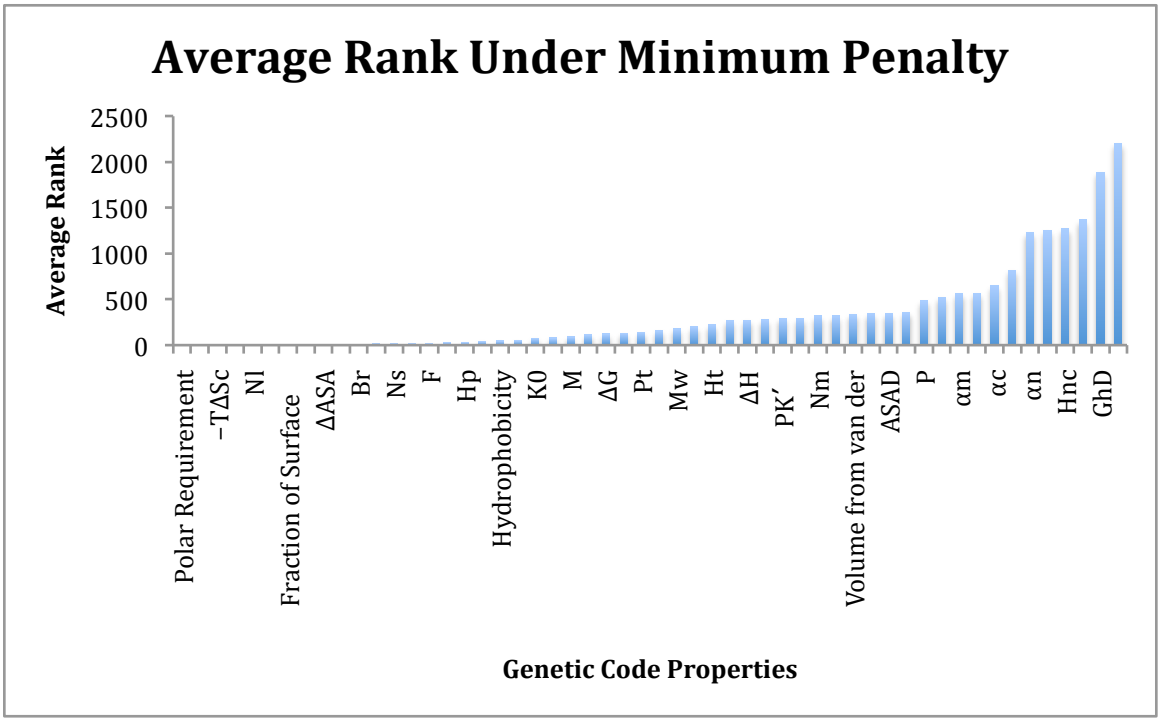
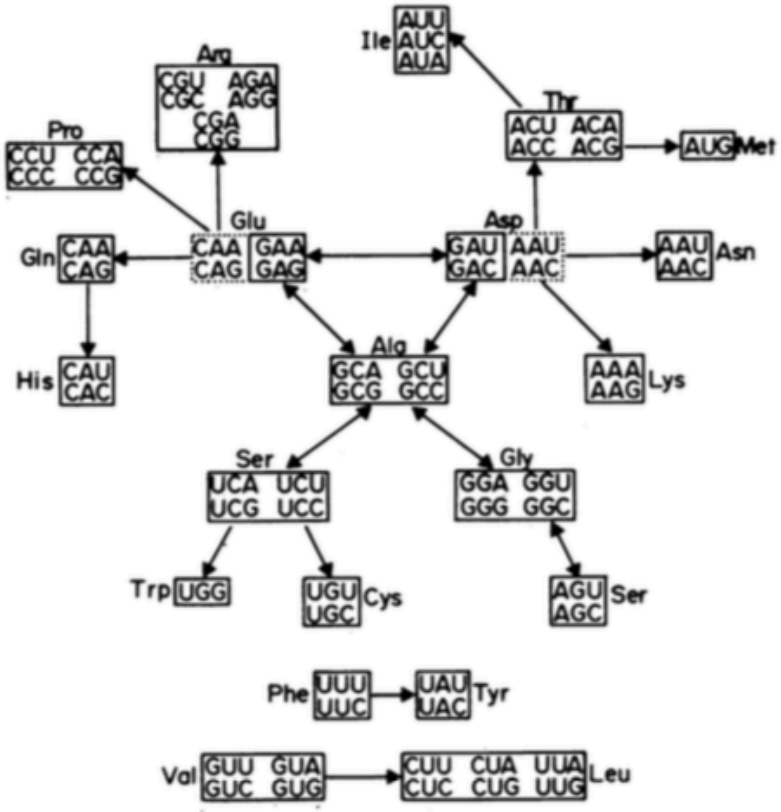
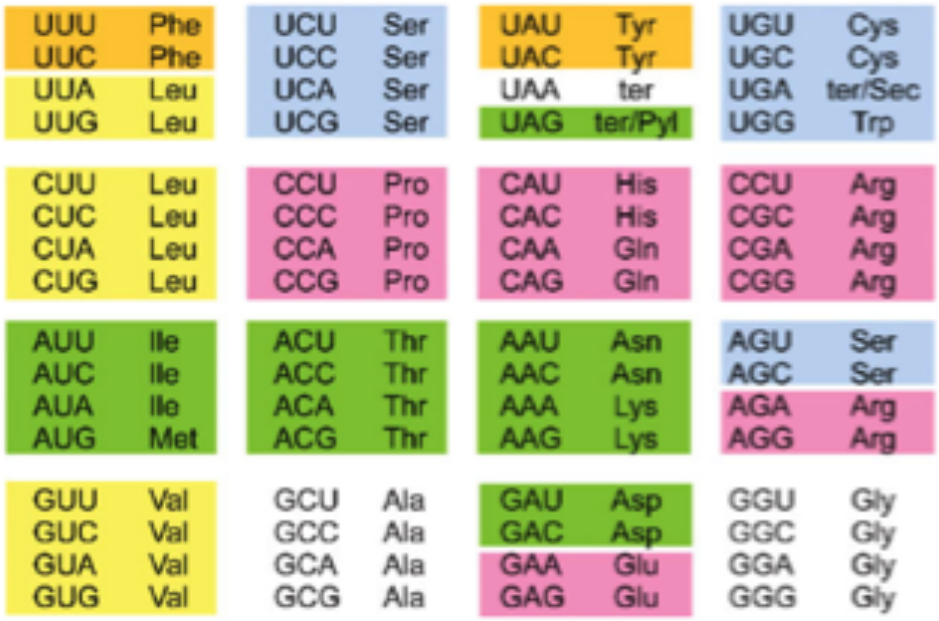


Figure 3: This bar graph represents the averaged ranks of the SGC among 10 sets of 2500 replicates for 54 different properties when minimum penalties are applied to stop codons.

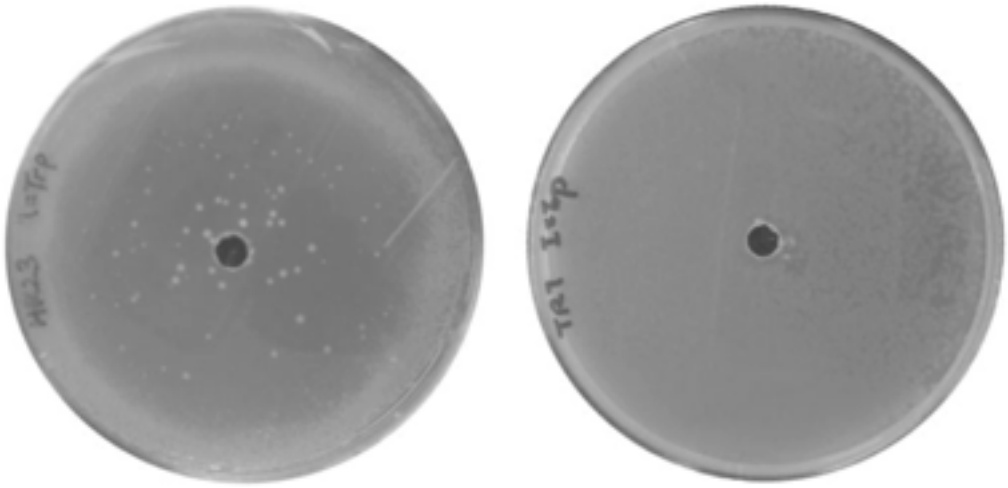
Wong 1975, Figure 1



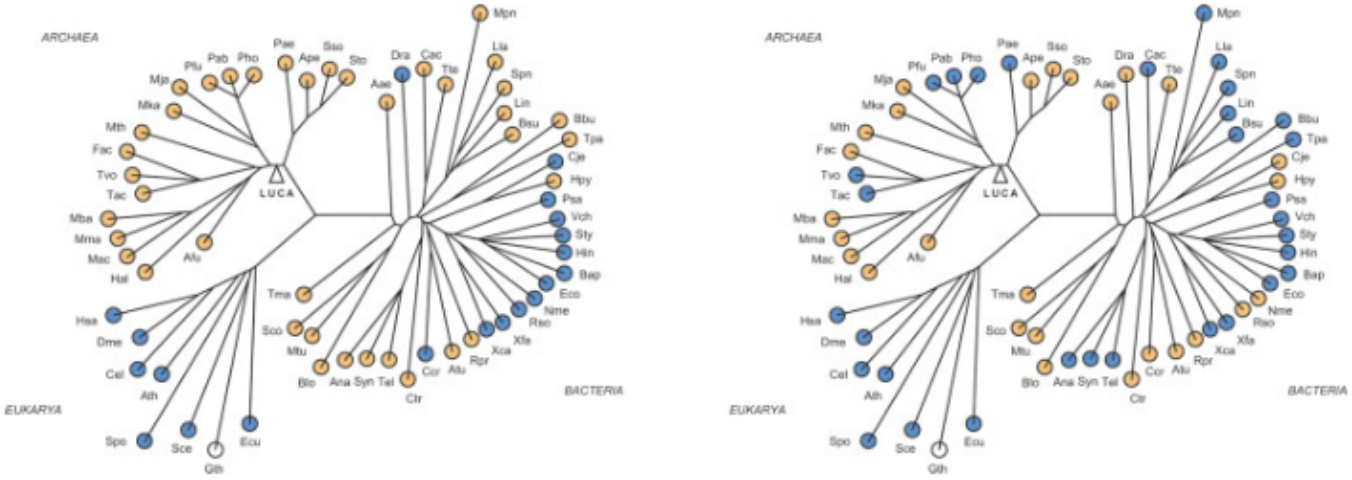
Wong 2005, Figure 1



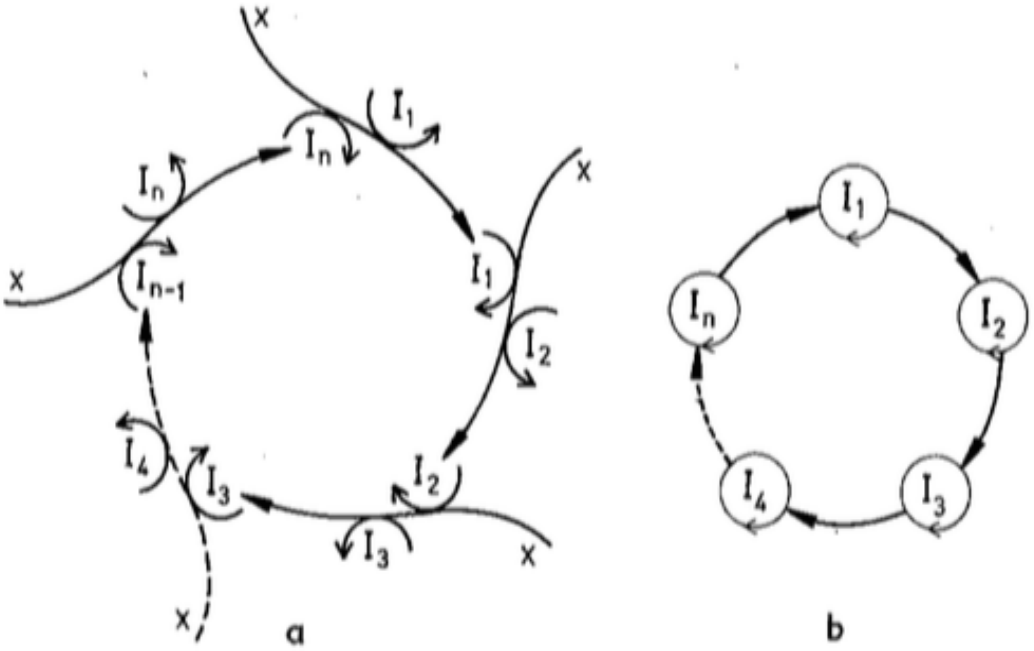
Wong 2005, Figure 2



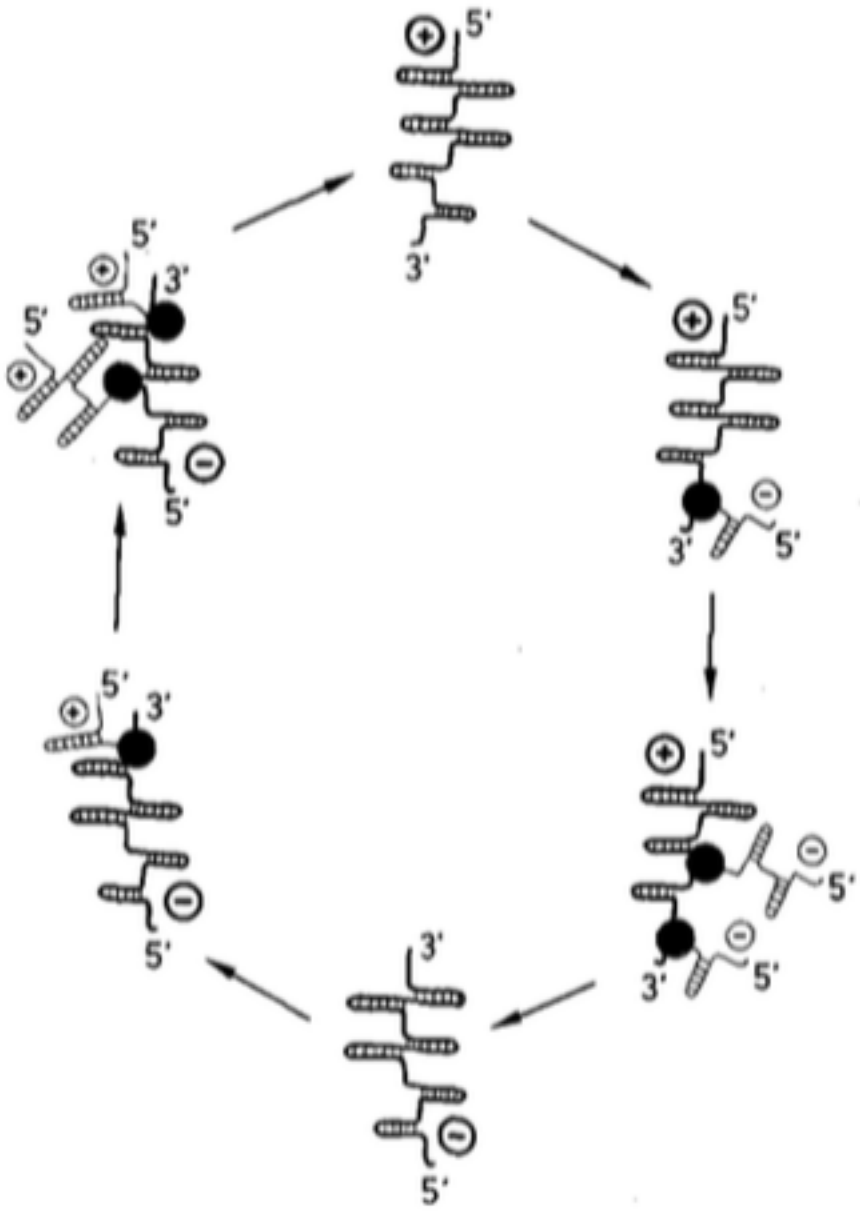
Wong 2005, Figure 3



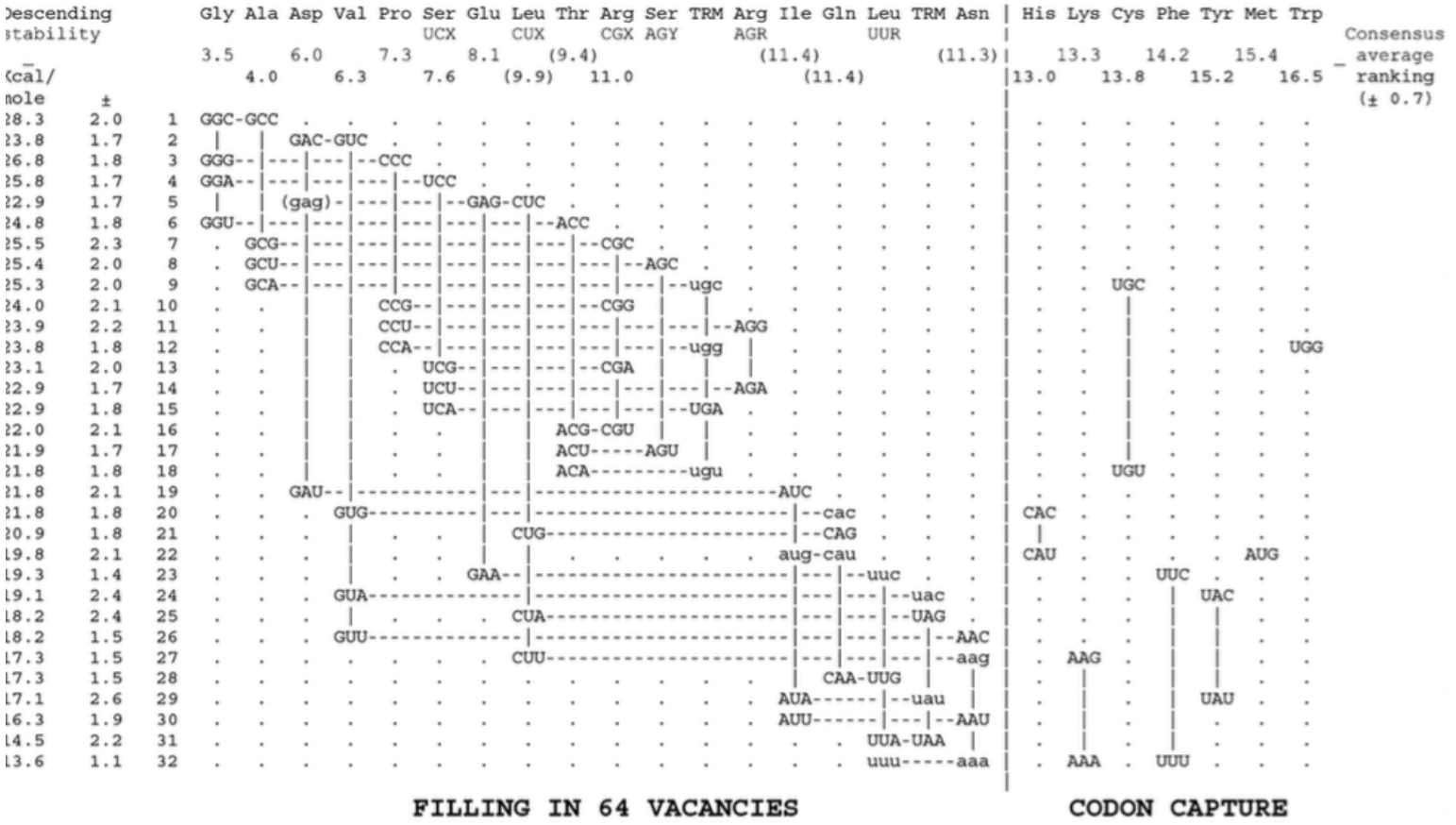
Eigen and Schuster 1977, Figure 7



Eigen and Schuster 1977, Figure 11



Trifonov 2004, Figure 1



Higgs 2009, Figure 1

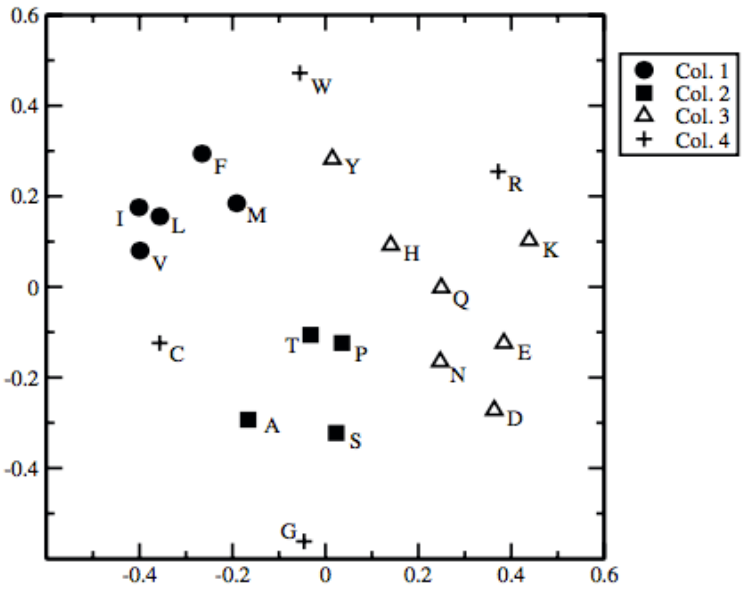


Figure 1

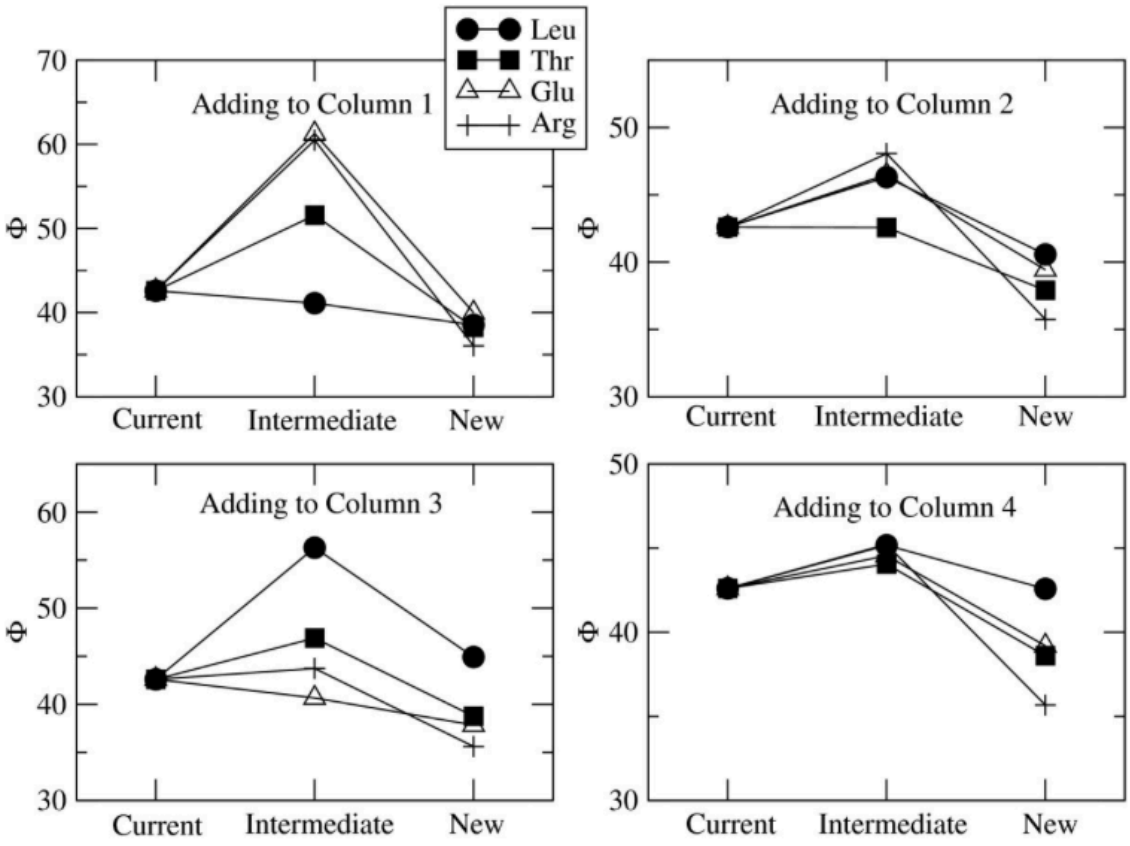
Higgs 2009, Figure 2

	U	C	A	G	
U	Val	Ala	Asp	Gly	U C
					A G
C					U C
					A G
A					U C
					A G
G					U C
					A G

Higgs 2009, Figure 3

<table>
<tr><th></th><th>U</th><th>C</th><th>A</th><th>G</th><th></th></tr>
<tr><td rowspan="2">U</td><td rowspan="4">Val
→ ?</td><td rowspan="4">Ala</td><td>Asp</td><td rowspan="2">Gly</td><td>U
C</td></tr>
<tr><td>Asp→ ?</td><td>A
G</td></tr>
<tr><td rowspan="2">C</td><td>Asp</td><td rowspan="2">Gly
→ ?</td><td>U
C</td></tr>
<tr><td>Asp→ ?</td><td>A
G</td></tr>
<tr><td rowspan="2">A</td><td rowspan="4">Val</td><td rowspan="2">Ala
→ ?</td><td>Asp</td><td rowspan="4">Gly</td><td>U
C</td></tr>
<tr><td>Asp→ ?</td><td>A
G</td></tr>
<tr><td rowspan="2">G</td><td rowspan="2">Ala</td><td>Asp</td><td>U
C</td></tr>
<tr><td>Asp→ ?</td><td>A
G</td></tr>
</table>

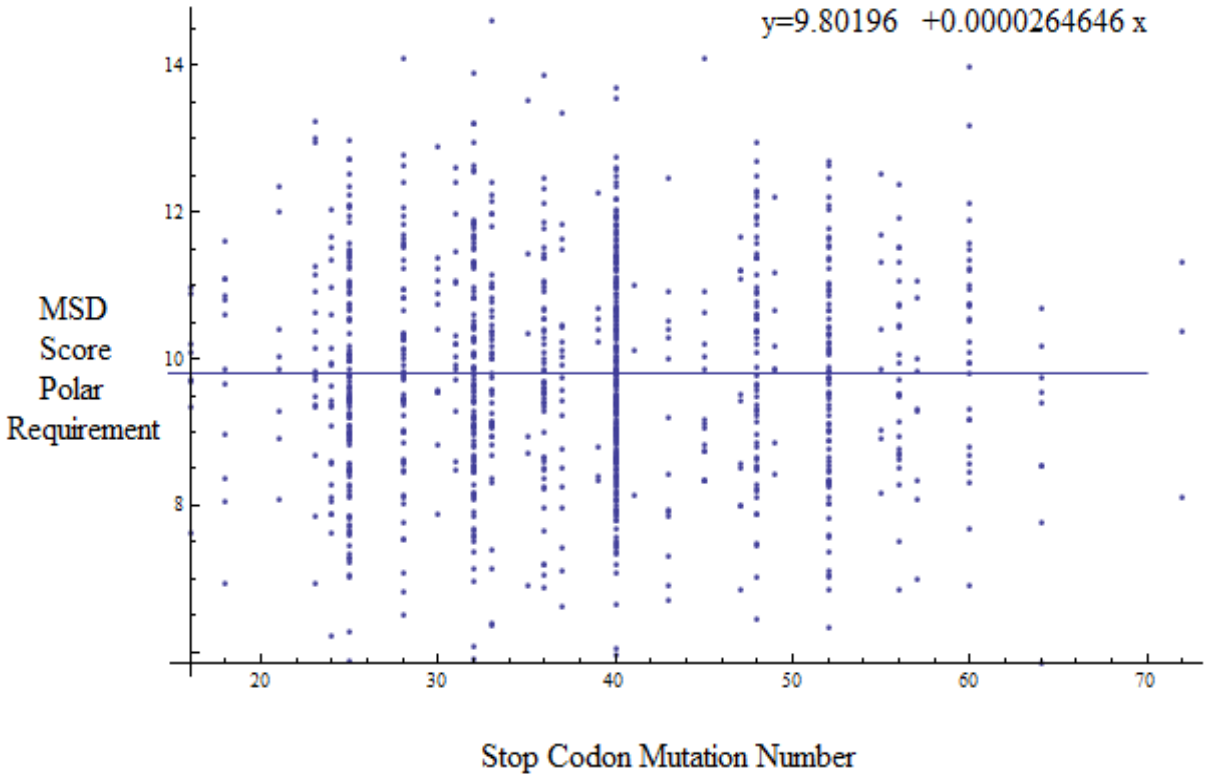
Higgs 2009, Figure 4



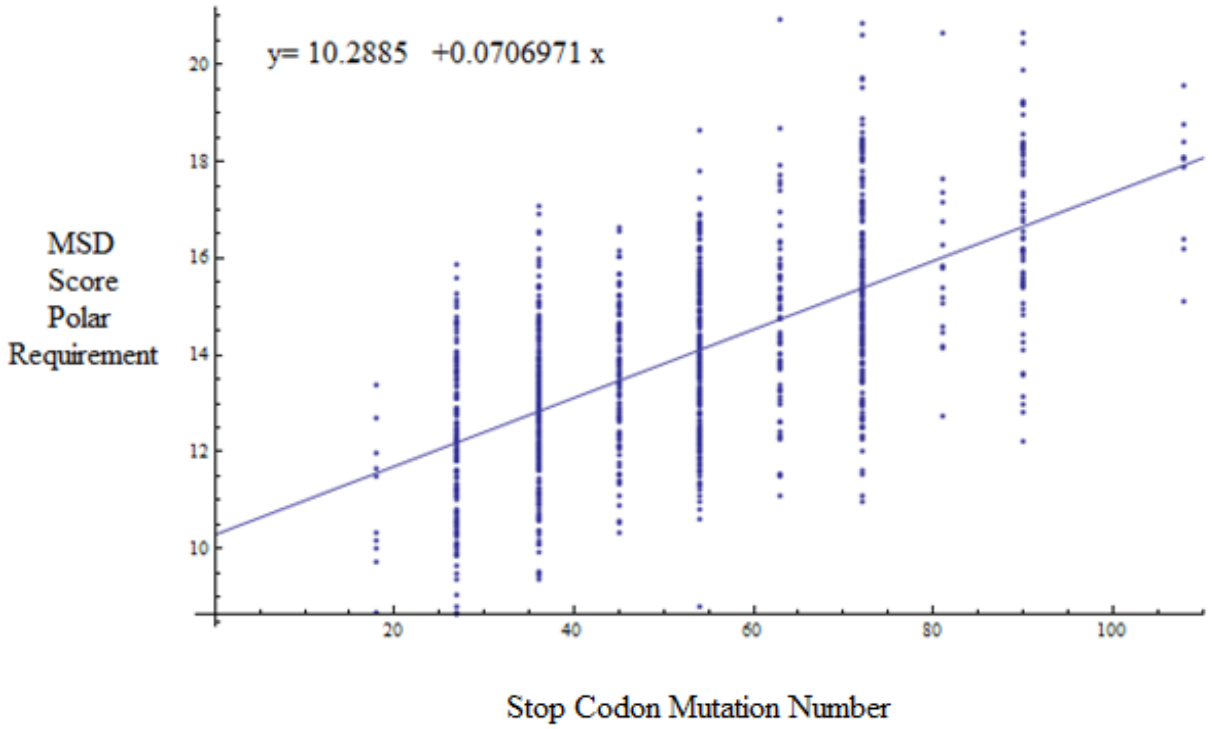
Higgs 2009, Figure 5

<table>
<tr><th></th><th>U</th><th>C</th><th>A</th><th>G</th><th></th></tr>
<tr><td rowspan="2">U</td><td rowspan="4">Leu</td><td rowspan="2">Ser</td><td>Asp</td><td rowspan="8">Gly</td><td>U
C</td></tr>
<tr><td>Glu</td><td>A
G</td></tr>
<tr><td rowspan="2">C</td><td rowspan="2">Pro</td><td>Asp</td><td>U
C</td></tr>
<tr><td>Glu</td><td>A
G</td></tr>
<tr><td rowspan="2">A</td><td rowspan="2">Ile</td><td rowspan="2">Thr</td><td>Asp</td><td>U
C</td></tr>
<tr><td>Glu</td><td>A
G</td></tr>
<tr><td rowspan="2">G</td><td rowspan="2">Val</td><td rowspan="2">Ala</td><td>Asp</td><td>U
C</td></tr>
<tr><td>Glu</td><td>A
G</td></tr>
</table>

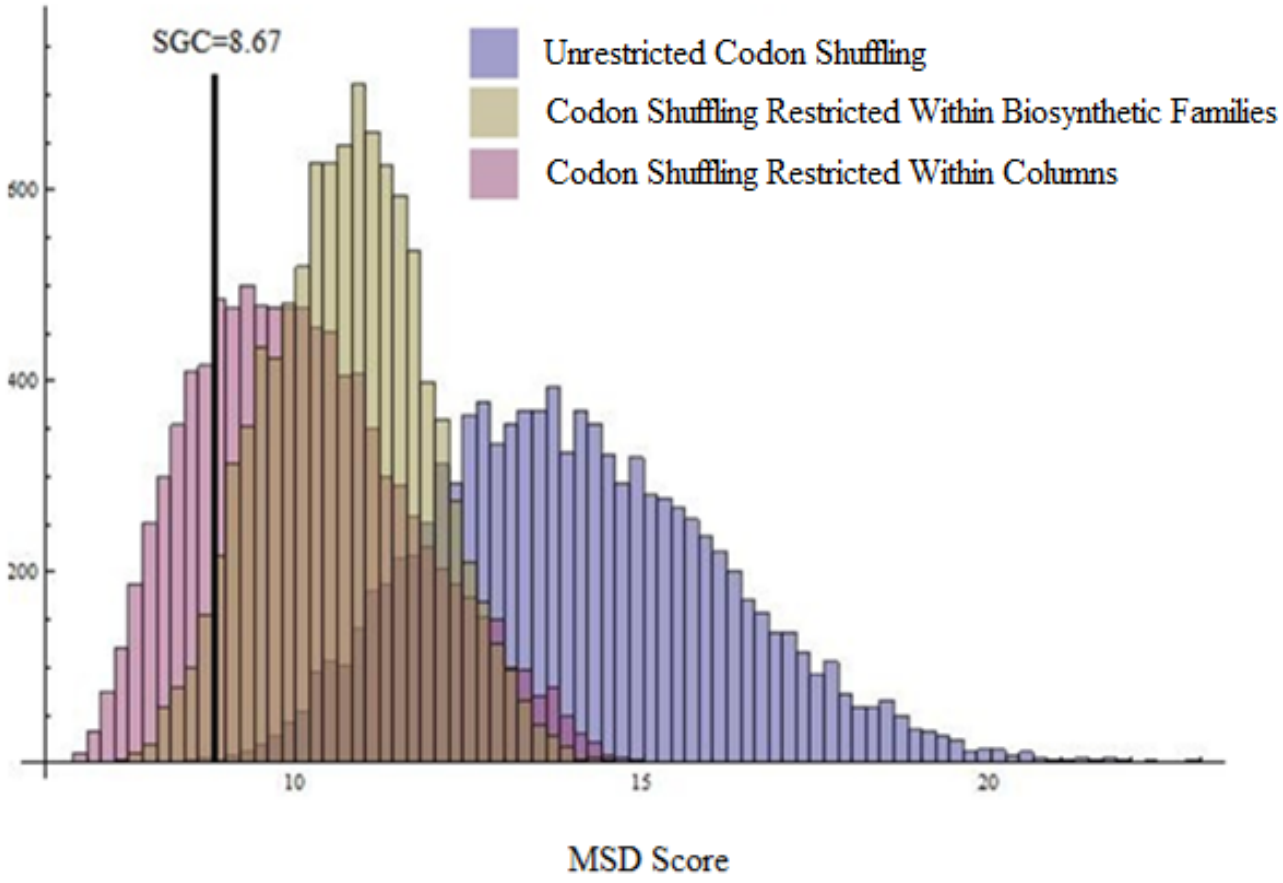
Degagne 2015, Figure 8



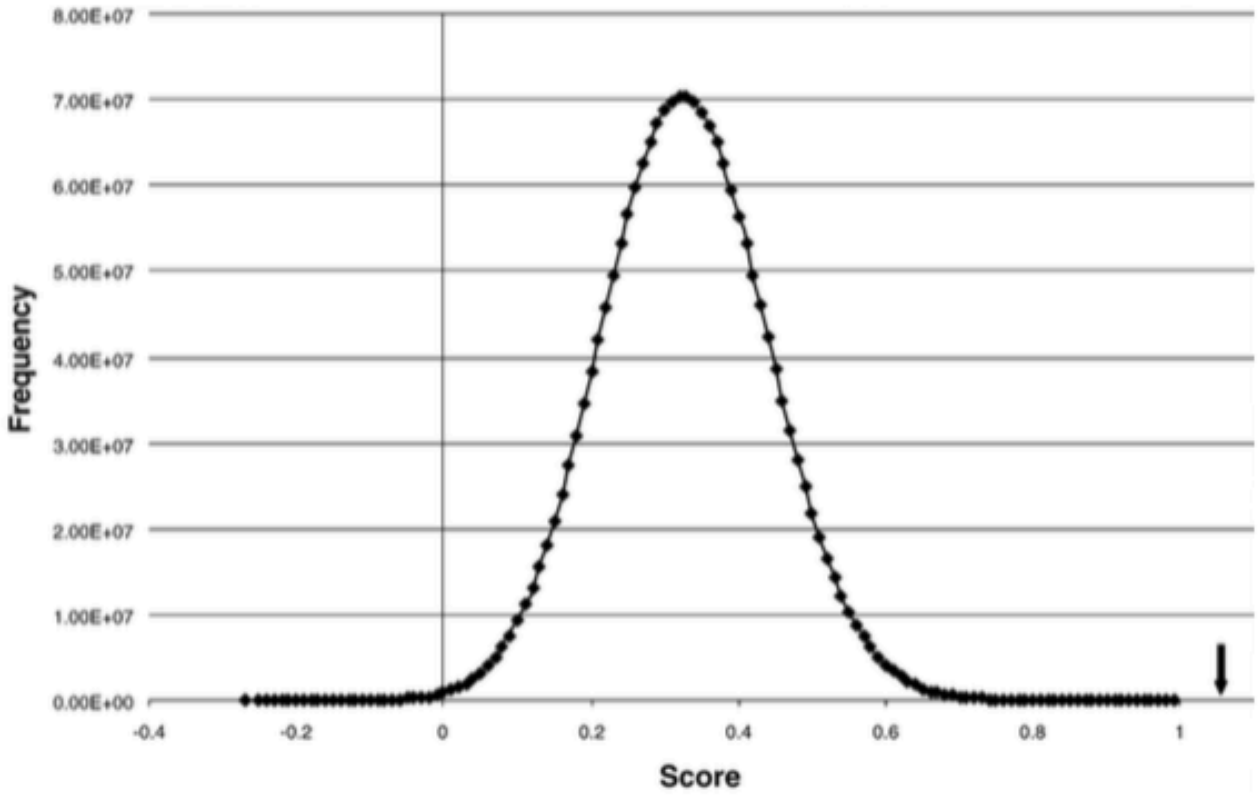
Degagne 2015, Figure 9



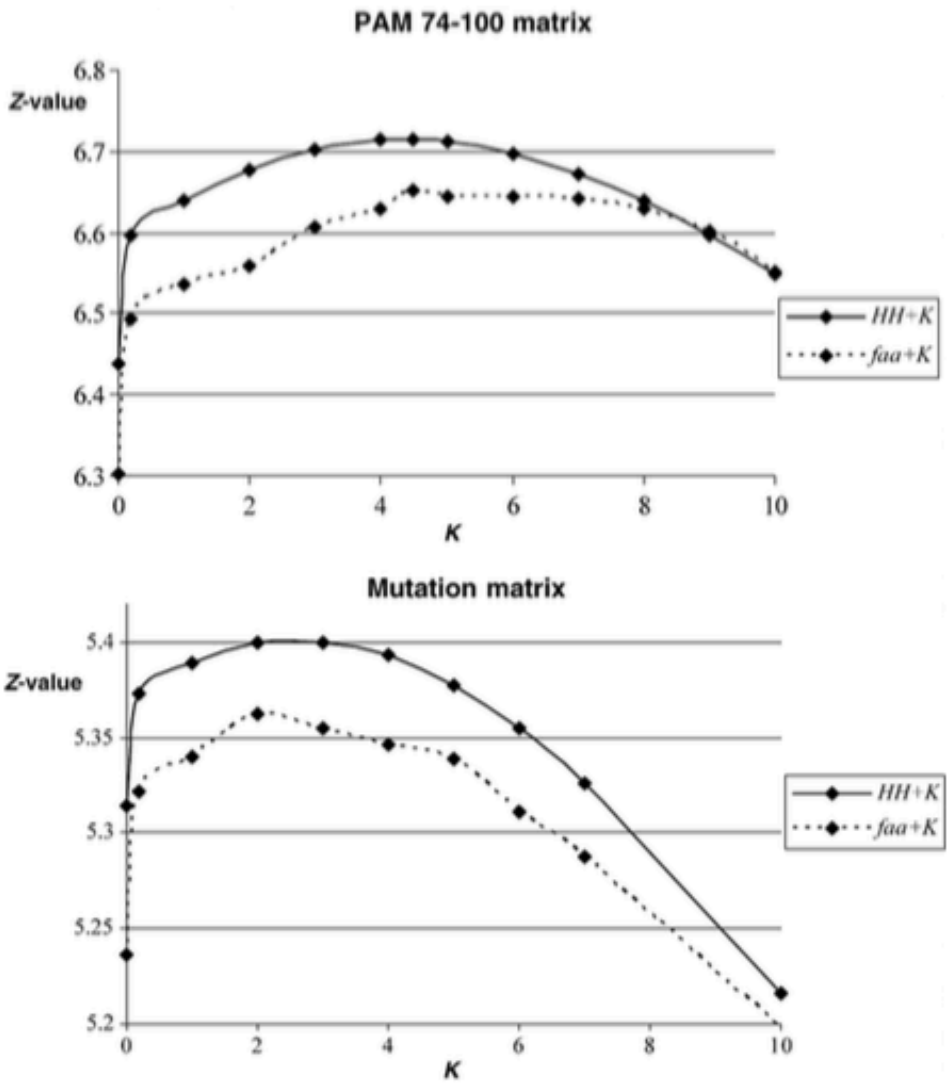
Degagne 2015, Figure 10



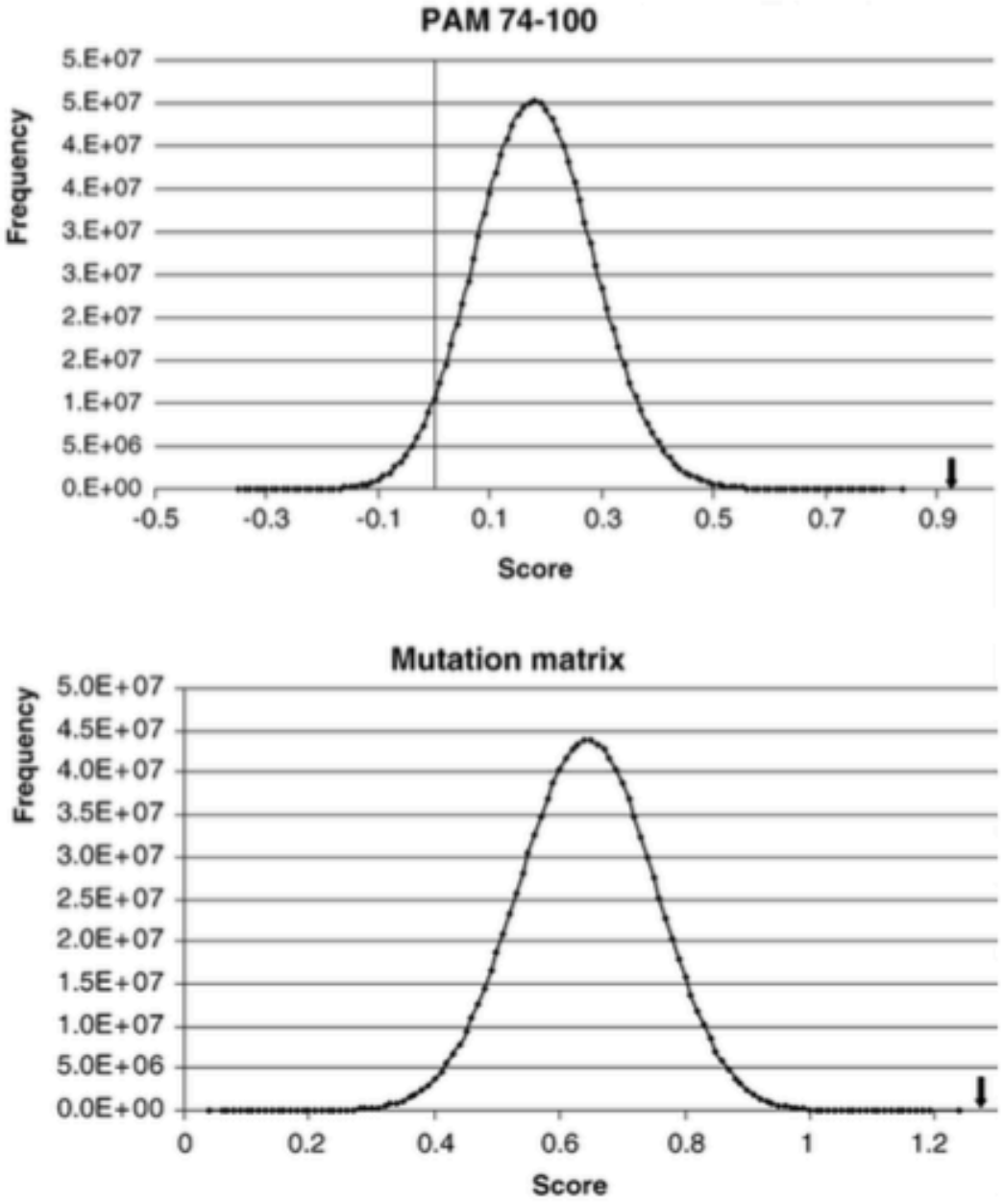
Goodarzi et al. 2004, Figure 2



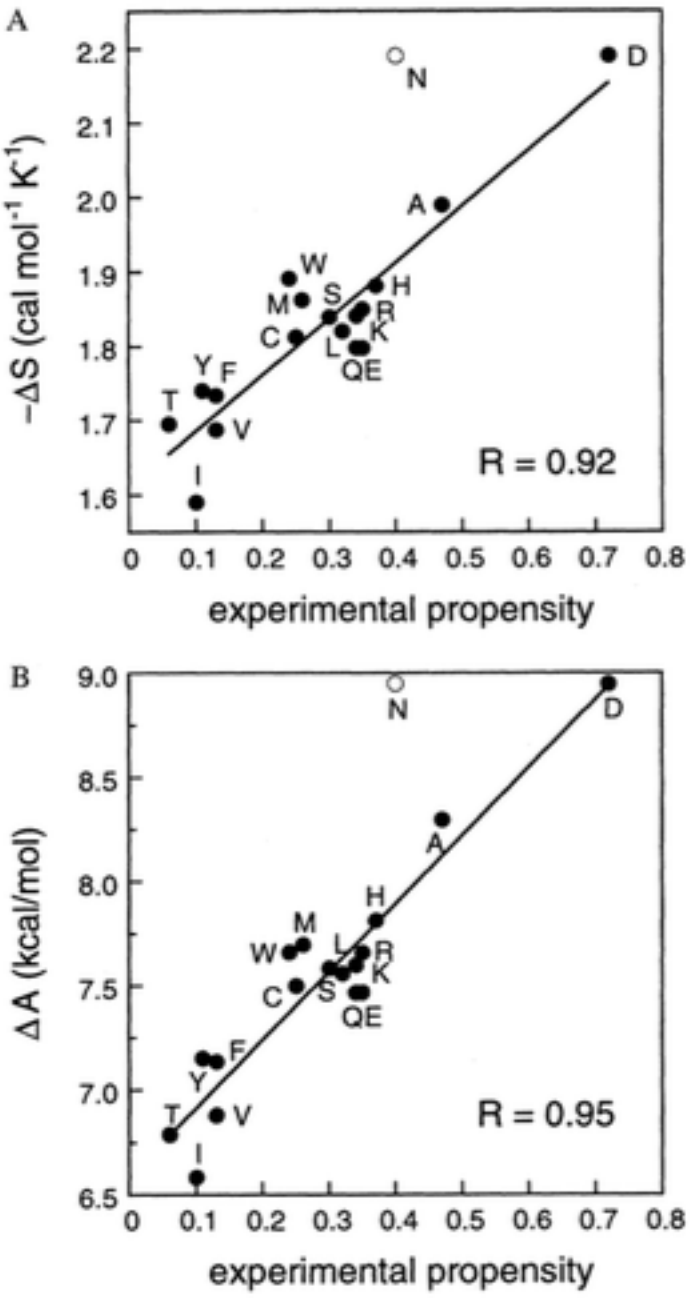
Goodarzi et al. 2004, Figure 3



Goodarzi et al. 2004, Figure 5



Street and Mayo 1999, Figure 1



Grimme 2004, Figure 4

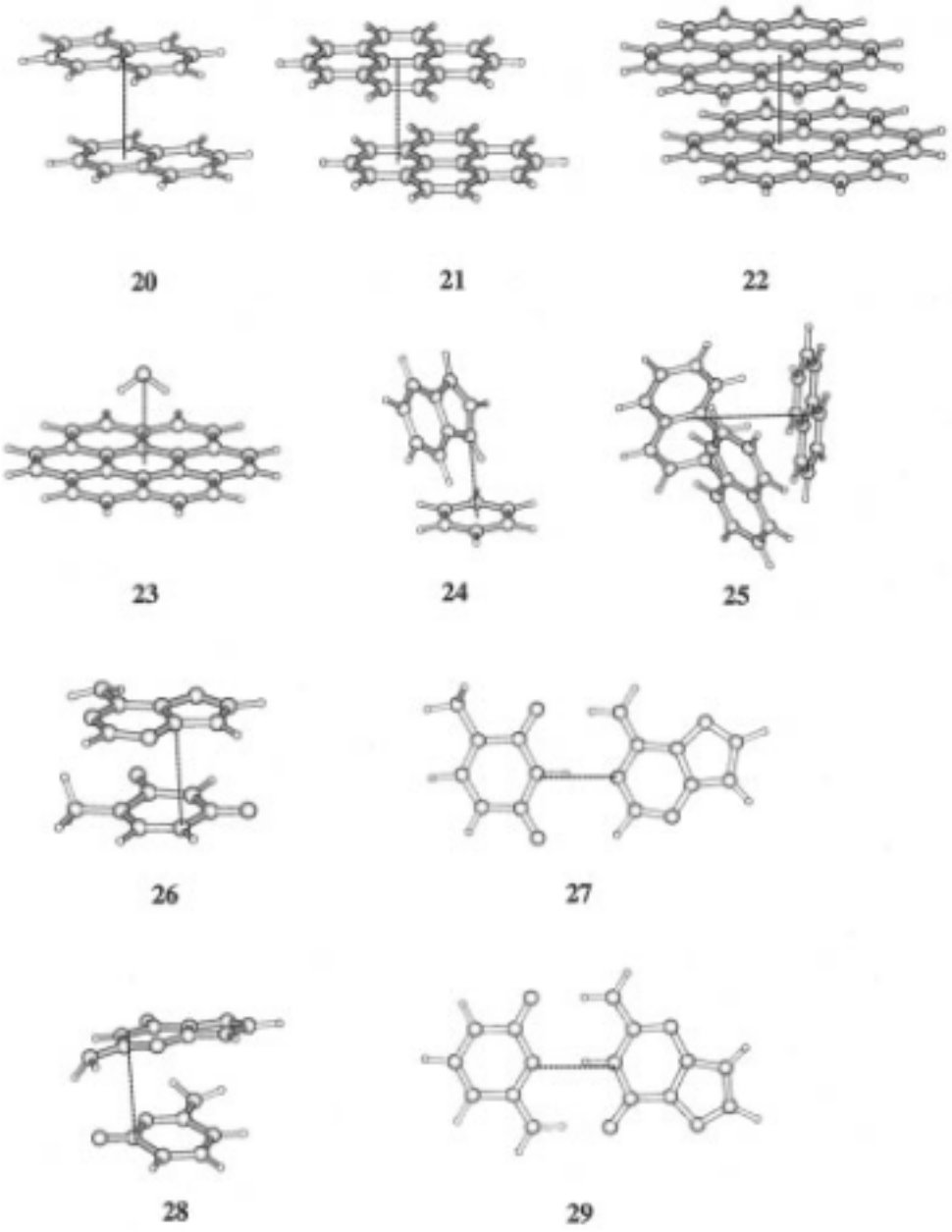


Table 1

Theory on the Genetic Code's Origin	Predicted Amino Acid Chronology – Phase 1	Predicted Amino Acid Chronology – Phase 2	Predicted Amino Acid Chronology – Phase 3
Coevolution Theory (1975)	-Glu -Asp -Ala -Ser -Gly	15 remaining amino acids	
First Principles Theory (2004)	-Gly -Ala -Asp -Val -Pro -Ser (UCX codon set) -Glu -Leu (CUX codon set) -Thr -Arg (CGX codon set)	-Ser (AGY codon set) -TERMINATION -Arg (AGR codon set) -Ile -Gln -Leu (UUR codon set) -TERMINATION -Asn	-His -Lys -Cys -Phe -Tyr -Met -Trp
Four-Column Theory (2009)	-Gly -Asp -Ala -Val	-Glu -Thr -Pro -Ser -Ile -Leu	10 Remaining amino acids

Table 2

1 = hydrophobicity	/
2 = polar requirement	/
3 = absolute entropy	/
4 = melting point	/
5 = compressibility	K^0
6 = thermodynamic transfer hydrophobicity	H_t
7 = surrounding hydrophobicity	Hp
8 = polarity	P
9 = isoelectric point	PH_i
10 = equilibrium constant with reference to the ionization property of cooh grou	PK'
11 = molecular weight	M_w
12 = bulkiness	B_l
13 = chromatographic index	R_f
14 = refractive index	M
15 = normalized consensus hydrophobicity	H_{nc}
16 = short and medium range non-bonded energy	E_{sm}
17 = long-range non bonded energy (london forces + electrostatic)	E_l
18 = total non-bonded energy	E_t
19 = alpha helix tendency	P_α
20 = b structure tendency	P_β
21 = turn tendency	P_t
22 = coil tendency	P_c
23 = helical contact area	C_a
24 = rms fluctuational displacement	F
25 = buriedness	B_r
26 = solvent accessible reduction ratio	R_a
27 = average number of surrounding residues	N_s
28 = power to be at the n terminal	α_n
29 = power to be at the c terminal	α_c
30 = power to be at the middle of an alpha helix	α_m
31 = partial-specific volum	V^0
32 = average medium contacts	N_m
33 = long range contacts (inter molecular stabilization)	N_l
34 = combined surrounding hydrophobicity	H_{gm}
35 = solvent accessible surface area for denatured proteins	ASA_D
36 = solvent accessible surface area for native proteins	ASA_N
37 = solvent accessible surface area for unfolding	ΔASA
38 = gibbs free energy change of hydration for unfolding	ΔG_h

39 = gibbs free energy change for denatured proteins	G_{hD}
40 = gibbs free energy change for native proteins	G_{hN}
41 = unfolding enthalpy change of hydration	ΔH_h
42 = unfolding entropy change of hydration	$-T\Delta S_h$
43 = unfolding hydration heat capacity change	ΔC_{ph}
44 = unfolding gibbs free energy	ΔG_c
45 = unfolding enthalpy	ΔH_c
46 = unfolding entropy change of hydration	$-T\Delta S_c$
47 = gibbs free energy change	ΔG
48 = unfolding enthalpy	ΔH
49 = unfolding entropy changes of the chain	$-T\Delta S$
50 = volume from van der Waals	/
51 = hydrophobicity scale	/
52 = hydrophobicity scale B	/
53 = surface area accessible to water	/
54 = fraction of surface area lost when folded	/

Table 3

Property	Standard Deviations Among Maximum Penalty Rankings	Standard Deviations Among Zero Penalty Rankings	Standard Deviations Among Minimum Penalty Rankings
Hydrophobicity	7.381659	5.349974	5.254628
Polar Requirement	1.032796	0.516398	0.316228
Absolute Entropy	23.30595	14.33565873	21.19853
Melting Point	801.7388	4.840799	54.09056
K^0	20.25504	8.066391	8.300602
H_t	14.87578	17.55911	13.83394
Hp	4.447221	4.6916	5.384133
P	27.56487	25.55517	15.44345
PH_i	11.74734	16.85757	27.09572
PK'	19.46621	14.23181	8.615877
M_w	13.49238	8.527081	16.18641
B_l	27.57293	13.80982	17.62227
R_f	1.75119	1.100505	1.032796
M	6.494442	6.003703	10.20022
H_{nc}	18.33909	30.72024	39.41009
E_{sm}	22.42048	6.040603	5.249339
E_l	0.421673	0.316228	0.316228
E_t	5.452828	3.098387	3.438346
P_α	15.73425	17.1153	9.82966
P_β	10.12148	7.439385	8.472832
P_t	14.27352	11.90518	4.903513
P_c	23.41154	16.95746	8.543353
C_a	536.8832	93.31672	71.58243
F	4.383048	5.526703	2.806738
B_r	3.645393	2.424413	2.877113
R_a	3.573047	2.065591	2.58414
N_s	5.270463	3.871549	3.852849
α_n	22.61784	22.02246	23.43691
α_c	29.9379	15.82719	14.9874
α_m	22.7303	16.17302	14.8357
V^0	20.07348	21.3947	16.74283
N_m	11.46056	15.45639	19.31062
N_l	2.282786	1.595131	1.523884
H_{gm}	1.563472	2.867442	6.684144
ASA_D	25.44799	12.28549	19.19751

ASA_N	22.57358	16.76769	11.35537
ΔASA	2.424413	2.836273	2.94392
ΔG_h	24.83814	22.79766	24.15137
G_{hD}	22.82472	22.69337	22.60801
G_{hN}	21.43621	14.19076	14.04319
ΔH_h	32.90795	22.46874	17.1127
$-T\Delta S_h$	13.58471	11.80442	8.232726
ΔC_{ph}	2.299758	1.969207	1.75119
ΔG_c	18.14907	14.42837	13.05927
ΔH_c	1.988858	3.093003	2.780887
$-T\Delta S_c$	1.37032	0.666667	0.483046
ΔG	10.88577	14.01943	9.131752
ΔH	15.08089	13.04863	11.46783
$-T\Delta S$	14.50249	18.27719	18.36331
Volume from van der Waals	224.338	72.48172	43.85633
Hydrophobicity Scale	3.802046	3.645393	4.794673
Hydrophobicity Scale B	23.84114	18.87532	19.01929
Surface Area Accessible to Water	33.49975	160.6144	10.78064
Fraction of Surface Area Lost When Folded	1.074968	1.712698	1.595131

Wong 1975, Table 2

TABLE 2. *Random probability of precursor–product codon contiguities*

Precursor–product	*a*	*b*	*n*	*x*	*P*
Ser–Trp	34	24	1	1	0.586
Ser–Cys	34	24	2	2	0.339
Val–Leu	24	36	6	6	0.00268
Thr–Ile	24	36	3	3	0.0591
Gln–His	14	48	2	2	0.0481
Phe–Tyr	14	48	2	2	0.0481
Glu–Gln	14	48	2	2	0.0481
Asp–Asn	14	48	2	2	0.0481

The parameters *a*, *b*, *n*, *x*, and *P* are defined by Eq. 1.

Wong 1975, Table 3

TABLE 3. *Stem sequence of the anticodon loop of some pairs of biosynthetically related transfer RNAs from Escherichia coli (19–21)*

tRNA pair	5′-arm.....3′-arm
Ser-1, Trp	C-C-G-G-U.....A-C-C-G-G
Thr, Ile-1	C-A-C-C-C.....G-G-G-U-G
Gln-2, His-1	C-Y-G-G-A.....Ψ-C-C-R-G
Val-2, Leu-2	C-Y-A-C-C.....G-G-U-R-G
Ala-1, Asp-1	C-C-U-G-C.....G-C-A-G-G

R = purine nucleoside; Y = pyrimidine nucleoside.

Wong 2005, Table1

Table 1. Comparison of phases 1 and 2 amino acids and amino acids produced by high energy proton irradiation of a carbon monoxide–nitrogen–water mixture

	Gly	Ala	Ser	Asp	Glu	Val	Leu	Ile	Pro	Thr	Phe	Tyr	Arg	His	Trp	Asn	Gln	Lys	Cys	Met
Phase of entry[a]	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
Irradiated synthesis[b]	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	n	n

[a]Phase 1–phase 2 entries into the code are as described (27), where Pro and Thr are regarded as marginal phase 1, and Phe, Tyr and Cys marginal phase 2 amino acids.
[b]Observed irradiated synthesis (51,52) is indicated by +, and lack of synthesis by 0. The comparison is not applicable to Cys and Met (n) because there was no sulfur in the irradiated synthesis.

Wong 2005, Table 2

Table 2. Encoding of amino acids by pretran synthesis[28,34,49,108]

Pretran synthesis	Nature of evidence	Organisms
Met-tRNA → fMet tRNA	Lack of fMetRS Presence of formylase	Bacteria
Glu-tRNA → Gln-tRNA	Lack of GlnRS Presence of amidotransferase	Archaea Bacteria
Asp-tRNA → Asn-tRNA	Lack of AsnRS Presence of amidotransferase	Archaea Bacteria
Ser-tRNA → Sec-tRNA	Lack of SecRS Elucidation of pretran pathway	All three domains

Trifonov 2004, Table I

Table I
Single-factor criteria of amino-acid chronology.

N1'. Criteria based on various evaluations of complexity of amino-acids. N1, N34, N35, N37 – as in (5), and N44 (6).
G, A, S, P, V, C, D, T, N, L, K, I, E, (MQ), H, R, F, Y, W

N2'. Criteria based on evolution of amino-acyl-tRNA synthetases. N2, N7 (5).
(AG), (DFHKNPST), (CEILMQRV), (WY)

N3'. Yields of amino acids in imitated primordial conditions. N3, N20, N21, N22 (5).
G, A, L, V, D, E, I, P, S, T, F, M, Y, K, (CHNQRW)

N4'. Criteria based on amino-acid compositions of various sets of presumably ancient proteins. N4, N26 (5), N41 (7), N45 (8), N47 (9), N48 (10), and N58 (11).
A, G, V, S, D, L, T, P, E, I, R, K, N, Q, H, F, M, Y, C, W

N5'. Criteria based on chemical inertness of amino acids. N5 (5), N42 (7), N49 (12).
G, A, V, S, (IL), D, (NQ), (FP), T, C, (EHKRWY), M

N6'. Stability of complementary interactions. N6, N39, N40 (5).
A, G, S, P, R, D, T, C, E, (VW), H, L, (MQ), I, N, Y, F, K

N8'. Stability of codon assignments. N8 (5), N52 (13).
(ADEFGHP), V, S, T, L, R, I, Q, (KNY), (MW), C

N11. GCU based theory (14).
A, (DGPSTV), E, (CFHIKLMNQRWY)

N12. RRY hypothesis of Crick (15).
(DGNS), (ACEFHIKLMPQRTVWY)

N13. RNY hypothesis (4).
(AG), (DINSTV), (CEFHKLMPQRWY)

N15. (excluded) Hypothesis of Ferreira (16).
(FGKLNP), (CDEHQRSTVW), (AIMY)

N17. (excluded) Early copolymerization code of Nelsestuen (17).
(DEFHIKLMSTVY), (ACGNPQRW)

N18. (excluded) Composition of proteinoids of Fox (18).
A, E, V, (GK), M, L, C, Y, (NQ), I, (DF), R, H, P, W, T, S

N24. Primordial code in tRNA sequence (19).
(ADGV), (CEFHIKLMNPQRSTWY)

N25. (excluded) Evolutionary distances between isoacceptor tRNAs (20).
Q, H, P, (LS), G, C, W, R, V, (DE), A, Y, T, (IM), F, (KN)

N27. (excluded) Match scores of BLOSSUM matrix (21).
(AILSV), (EKMQRT), (DFGN), (PY), H, C, W

N31. (excluded) Algebraic model of Hornos and Hornos (22).
(CDFSV), (EKLRY), (HP), (AGIMNQTW)

N32. Composition of translated Urgen (23, 24).
V, (AGP), (ENRT), (LQS), (CDFHIKMYW)

N33. Murchison meteorite (25).
(AG), (DEPV), (CFHIKLMNQRSTWY)

N38. Minimal alphabet for folding (26).
(AGEIK), (CDFHLMNPQRSTVWY)

N43. Mutational stability of codon assignments (27). More stable assignments – earlier.
(AGPTV), L, R, S, I, (CDEFHKNQY), (MW)

N46. (excluded) Complementary circular code (28).
(ADEFGIKLNQTVY), (CHMPRSW)

N51. (excluded) Metabolic cost (H. Akashi, personal communication). Costly amino acids – late.
D, N, T, E, Q, K, P, I, M, S, G, A, R, V, L, C, H, Y, F, W

N56. Composition of an urancestral tRNA gene (29).
S, G, P, (LQV), H, (EINRTW), A, (CDFKMY)

N59. Arginine first (30).
R, (ACDEFGHIKLMNPQSTVWY)

Trifonov 2004, Table II

Table II
Consensus temporal order of amino acids (single-factor criteria)

Amino acids of Miller		Average rank (± 0.7)	Order	Codon capture cases
+	G	2.8	1	
+	A	3.9	2	
+	V	6.5	3	
+	S	7.1	4	
+	P	7.4	5	
+	D	7.7	6	
+	T	9.0	7	
+	E	9.9	8	
+	L	10.3	9	(+)
+	I	10.9	10	(+)
	N	11.2	11	
	R	11.7	12	
	H	12.7	13	+
	Q	12.8	14	+
	K	13.2	15	
	F	13.2	16	+
	C	13.9	17	+
	M	15.0	18	+
	W	15.3	19	+
	Y	15.3	20	+

Trifonov 2004, Table III

Table III

Multi-factor criteria of amino-acid chronology.

N9. Jukes' theory of the origin of the code (32).
(ADEGHLPQRV), (CFIKNSTY), (MW)

N10. Coevolution theory of Wong (34-36).
(ADEGS), V, (PT), (IL), F, C, Y, (KR), (NQ), H, (MW)

N14. Hypothesis of Hartman (38).
G, P, A, R, (DENQST), (HK), C, (FILVY), (MW)

N16. Prebiotic physicochemical code of Altshtein-Efimov (39).
(ADEGKRSTV), (CFHILMNPQWY)

N19. Coevolution theory of Dillon (33).
G, A, D, V, E, Q, (HLPR), N, T, (IS), (KM), F, (CY), W

N23. Coevolution theory of Wächtershäuser (37).
(DE), (ACGNPQST), (ILMV), (FHKRWY)

N28. (excluded) A/U start theory (40).
(FIKLMNY), (CDEHQRSTVW), (AGP)

N29. N-fixing amino acids first, Davis (41).
(DENQ), (APSV), (CG), T, (ILM), R, K, (FY), H, W

N30. GNN codons first, Taylor and Coates (42).
(ADEGV), (CFHIKLMNPQRSTWY)

N36. Jimenez-Montano (43).
(ADGV), (LPR), (CIKQST), (EFHMNWY)

N50. SNS code of Ikehara (44).
(ADGV), E, (HLPQR), (CFIKMNSTWY)

N53. Cavalier-Smith (45).
(AILPV), (DEGST), (CFHNY), (KMQRW)

N54. (excluded) Self-referential code (46).
(GPS), (DEFKLN), (AHQRV), (CIMTWY)

N55. (excluded) Synthesis with hydropathicity (47).
G, S, D, N, K, (AF), (HT), E, Q, L, (PV), R, (CW), (IM), Y

N57. Maslov, GGG start (48).
G, (DS), A, (LPV), (EFHIKMNQR), (CTWY)

N60. Three stages of Baumann and Oro (49).
(ADEGILPQRSTV), (KN), (CFHY), (MW)

Trifonov 2004, Table IV

Table IV
Consensus temporal order of amino acids (multi-factor criteria).

Amino acids of Miller		Average rank (± 0.7)	Order	Codon capture cases
+	A	4.1	1	
+	G	4.2	2	
+	D	4.2	3	
+	V	6.1	4	
+	E	6.3	5	
+	P	7.2	6	
+	S	8.0	7	
+	L	9.5	8	(+)
+	T	9.8	9	
	Q	9.9	10	(+)
	R	10.2	11	
	N	11.4	12	
+	I	11.9	13	(+)
	H	13.2	14	+
	K	13.4	15	
	C	13.8	16	+
	F	15.1	17	+
	Y	15.2	18	+
	M	15.9	19	+
	W	17.7	20	+

Trifonov 2004, Table V

Table V
Consensus temporal order of amino acids (final).

Amino acids of Miller		Average rank (± 0.7)	Order	Codon capture cases
+	G	3.5	1	
+	A	4.0	2	
+	D	6.0	3	
+	V	6.3	4	
+	P	7.3	5	
+	S	7.6	6	
+	E	8.1	7	
+	T	9.4	8	
+	L	9.9	9	(+)
	R	11.0	10	
	N	11.3	11	
+	I	11.4	12	(+)
	Q	11.4	13	(+)
	H	13.0	14	+
	K	13.3	15	
	C	13.8	16	+
	F	14.2	17	+
	Y	15.2	18	+
	M	15.4	19	+
	W	16.5	20	+

Higgs 2009, Table 2

Table 2: Barriers (δΦ) and net changes in cost (ΔΦ) for addition of amino acids to an 8-codon block in column 1

	$\varepsilon = 0.05$		$\varepsilon = 0$	
	δΦ	ΔΦ	δΦ	ΔΦ
Leu	-1.49	-4.07	-1.56	-4.54
Ile	-0.99	-3.26	-1.07	-3.69
Val	0.00	0.00	0.00	0.00
Phe	1.09	-4.02	1.07	-4.84
Met	2.00	-3.83	2.11	-4.63
Cys	6.60	0.01	6.92	-0.67
Tyr	6.68	-4.66	6.94	-6.17
Trp	8.33	-1.72	8.50	-3.21
Thr	8.97	-4.38	9.55	-5.98
Pro	11.20	-3.69	11.84	-5.70
His	11.51	-4.67	12.05	-6.79
Ala	11.67	1.19	12.33	0.00
Gln	14.49	-6.81	15.16	-9.69
Ser	15.25	-1.36	16.05	-3.36
Asn	16.73	-3.65	17.51	-6.41
Arg	17.90	-6.56	18.52	-9.69
Glu	18.68	-2.61	19.44	-5.71
Lys	19.25	-6.51	19.96	-9.71
Asp	20.70	2.50	21.54	0.00
Gly	21.60	2.30	22.48	0.00

Higgs 2009, Table 3

Table 3: Barriers (δΦ) for addition of amino acids to codon blocks in columns 2, 3 and 4 in the positions indicated in Figure 3

Col 2 (4 codons)			Col 3 (4 × 2 codons)			Col 4 (4 codons)		
	ε = 0.05	ε = 0		ε = 0.05	ε = 0		ε = 0.05	ε = 0
Thr	-0.03	0.01	Glu	-1.94	-2.02	Gly	0.00	0.00
Ala	0.00	0.00	Gln	-1.89	-1.93	Ser	0.77	0.92
Ser	0.10	0.06	Asn	-1.65	-1.67	Ala	0.98	1.09
Pro	0.53	0.57	Lys	-0.66	-0.73	Asn	1.44	1.69
Asn	2.10	2.13	Asp	0.00	0.00	Thr	1.45	1.65
Cys	2.27	2.41	Arg	1.13	1.11	Pro	1.62	1.83
Gln	2.63	2.72	His	1.22	1.31	Asp	1.65	1.87
Gly	2.65	2.61	Pro	3.81	4.03	Cys	1.75	1.84
His	2.75	2.87	Thr	4.28	4.54	Gln	1.85	2.13
Met	3.23	3.47	Ser	5.42	5.73	His	1.86	2.10
Val	3.48	3.74	Tyr	6.28	6.53	Glu	1.95	2.21
Asp	3.62	3.65	Ala	9.33	9.80	Met	2.29	2.45
Leu	3.71	4.00	Met	9.89	10.32	Lys	2.42	2.68
Glu	3.89	3.96	Trp	11.65	12.04	Val	2.46	2.58
Tyr	3.97	4.18	Gly	12.10	12.64	Tyr	2.54	2.75
Ile	4.21	4.51	Phe	12.94	13.45	Leu	2.57	2.70
Phe	4.59	4.88	Cys	13.43	14.00	Arg	2.61	2.85
Lys	4.79	4.89	Leu	13.68	14.25	Ile	2.69	2.80
Arg	5.46	5.59	Val	14.26	14.85	Phe	2.78	2.91
Trp	6.20	6.46	Ile	15.15	15.76	Trp	3.25	3.39

Goodarzi et al. 2004, Table 2

Table 2
Precursor-product pairs, indicated by Ronneberg et al. (2000)

Precursor	Product	Precursor	Product
Ser	Trp	Asp	Asn
Ser	Cys	Asp	Met
Phe	Tyr	Asp	Lys
Thr	Ile	Glu	Pro
Gln	His	Asp	Arg
Glu	Gln	Asp	Thr

These are self consistent and biologically relevant set of precursors as the closest direct antecedents to their product amino acid in an energetically favorable non-transamination pathway.

Grimme 2004, Table 3

Complex		$-\Delta E_{DFT}$	DFT-D-BLYP		Reference		
			R	$-\Delta E$	R	$-\Delta E$	Method[a]
			Hydrogen bonded complexes				
1	$(NH_3)_2$ (C_{2h})	2.17	327	3.18 (3.59)	317	3.0	MP2/MP4[b]
2	$(H_2O)_2$ (C_s)	4.53	293	5.42 (5.98)	292	4.8	MP2[c]/CCSD(T)[d]
3	$(HF)_2$ (C_s)	4.48	276	4.96 (5.48)	275	4.4	MP2[c]/CCSD(T)[d]
4	$(HCOOH)_2$ (C_{2h})	14.06	268	16.09 (16.60)	266	13.9	MP2[c]/CCSD(T)[d]
			Non-aromatic complexes				
5	$(Ne)_2$ ($D_{\infty h}$)	−0.11	311	0.05 (0.20)	309	0.08	exp.[e]
6	$(CH_4)_2$ (D_{3d})	−0.80	363	0.31 (0.32)	360	0.5	MP2[d]
7	$CH_3F \cdot CH_4$ (C_s)	−0.43	263	0.98 (1.17)	261	0.7	MP2[f]
8	$CH_3F \cdot$ ethine (C_{3v})	1.21	230	1.62 (1.76)	219	1.7	MP2[f]
9	$(ethene)_2$ (D_{2d})	−1.22	373	1.36 (1.41)	380	1.3	MP2/CCSD(T)[d]
10	ethene · ethine (C_{2v})	0.38	389	1.35 (1.45)	382	1.52	MP2[c]
11	$(N_2)_2$ (D_{2d})	−0.24	368	0.29 (0.19)	341	0.33	MP2[c]
12	$(F_2)_2$ (D_{2d})	−0.40	304	0.27 (0.34)	306	0.27	MP2[c]
			Benzene complexes				
13	benzene · CH_4 (C_{3v})	−1.00	381	0.90 (0.96)	362	1.6	MP2[c]
14	benzene · NH_3 (C_s)	0.25	362	1.78 (2.16)	345	2.4	MP2[c]
15	benzene · H_2O (C_{2v})	0.70	333	3.13 (3.73)	321	3.9	MP2[g]
16	benzene · Ne (C_{6v})	−0.41	341	0.24 (0.56)	330	0.43	exp.[h]
			Benzene dimers				
17	$(benzene)_2$ (D_{6h})	−3.11	390	1.04 (1.21)	370	1.8	MP2/CCSD(T)[i]
18	$(benzene)_2$ (T, C_{2v})	−1.08	506	2.03 (2.20)	490	2.7	MP2/CCSD(T)[i]
19	$(benzene)_2$ (PD, C_{2h})	−3.39	352	2.00 (2.18)	340	2.8	MP2/CCSD(T)[i]
			Complexes of larger aromatic molecules				
20	$(naphthalene)_2$ (C_i)	−6.31	345	5.34 (5.58)	350	6.2	MP2/CCSD(T)[j]
21	$(pyrene)_2$ (C_{2h})	−11.36	343	11.82 (12.35)	353	13.1	exp.[k]/MP2[l]
22	$(coronene)_2$ (C_i)	−17.52	340	21.56 (22.21)	—	—	
23	coronene · H_2O (C_{2v})	0.10	330	3.20 (4.09)	339[m]	4.5	MP2[n]
24	benzene · indole (C_1)	−0.20	334	4.78 (5.15)	316	5.9	MP2/exp.[o]
25	$(naphthalene)_3$ (C_{3h})	−8.09	486	13.22 (13.82)	493	8.7	exp.[p]
			DNA base pairs				
26	A · T (S, C_1)	−2.36	344	10.47 (11.27)	—	11.6	MP2/CCSD(T)[q]
27	A · T (WC, C_s)	10.91	284	15.01 (15.54)	—	15.4	MP2/CCSD(T)[q]
28	G · C (S, C_1)	5.16	317	15.30 (16.53)	—	16.9	MP2/CCSD(T)[q]
29	G · C (WC, C_s)	23.56	293	28.10 (28.82)	—	28.8	MP2/CCSD(T)[q]