A Computational Analysis of the Structure of the Genetic Code

A Computational Analysis of the Structure of the Genetic Code

by

Christopher Degagne

A Thesis Submitted to the School of Graduate Studies
In Partial Fulfillment of the Requirements for the Degree
Master of Science

McMaster University
August 2015

Master of Science (2015) McMaster University
Department of Biology

Title: A Computational Analysis of the Structure of the Genetic Code

Author: Christopher Degagne, H.BSc.

Supervisor: Dr. Jonathon Stone

Number of Pages: 72

Abstract

The standard genetic code (SGC) is the cipher used by nearly all organisms to transcribe information stored in DNA and translate it into its amino acid counterparts. Since the early 1960s, researchers have observed that the SGC is structured so that similar codons encode amino acids with similar physiochemical properties. This structure has been hypothesized to buffer the SGC against transcription or translational error because single nucleotide mutations usually either are silent or impart minimal effect on the containing protein. We herein briefly review different theories for the origin of that structure. We also briefly review different computational experiments designed to quantify buffering capacity for the SGC.

We report on computational Monte Carlo simulations that we performed using a computer program that we developed, AGCT. In the simulations, the SGC was ranked against other, hypothetical genetic codes (HGC) for its ability to minimize physiochemical distances between amino acids encoded by codons separated by single nucleotide mutations. We analyzed unappreciated structural aspects and neglected properties in the SGC. We found that error measure type affected SGC ranking. We also found that altering stop codon positions had no effect on SGC ranking, but including stop codons in error calculations improved SGC ranking. We analyzed 49 properties individually and identified conserved properties. Among these, we found that long-range non-bonded energy is more conserved than is polar requirement, which previously was considered to be the most conserved property in the SGC. We also analyzed properties in combinations. We hypothesized that the SGC is organized as a compromise among multiple properties.

Finally, we used AGCT to test whether different theories on the origin of the SGC could explain more convincingly the buffering capacity in the SGC. We found that, without accounting

for transition/transversion biases, the SGC ranking was modest enough under constraints imposed by the coevolution and four column theories that it could be explained due to constraints associated with either theory (or both theories); however, when transition/transversion biases were included, only the four column theory returned a SGC ranking modest enough that it could be explained due to constraints associated with that theory.

Acknowledgements

I would like to acknowledge the support and guidance of both Dr. Jonathon Stone and Tarushika Vasanthan (PhD. Candidate) whose tireless support and guidance has made this possible. Without Dr. Stone's interest and advocacy, I know that I wouldn't be where I am today. May his grants be generous, may his publications proliferate, and may he one day be repaid for the patience and generosity that he has shown so many of his students. I would also like to thank Taru for her support, guidance, and encouragement. I has been a privilege working alongside her.

I would also like to thank my parents, Paul and Elizabeth Degagne, whose material support and encouragement, and general belief in my competence has not only encouraged me to achieve, but also given me the tools to do so. I am indebted to them in more ways then I'll ever know.

Thirdly, I would like to thank the Origins Institute, as well as NSERC, for funding my project.

Finally, I would likely to thank the many friends I have made at McMaster. In many ways they have become like family. Their support, trust, and confidence has inspired me to reach higher and try harder than ever before.

Table of Contents

Chapter 3: TESTING THEORIES OF THE STRUCTURE OF THE GENETIC CODE

List of Figures

List of Tables

List of Abbreviations

DNA: deoxyribonucleic acid
RNA: ribonucleic acid
mRNA: messenger ribonucleic acid
tRNA: transfer ribonucleic acid
SGC: standard genetic code
HGC: hypothetical genetic code
PGC: primordial genetic code
SD: squared difference
AD: absolute difference

Preface

The following body of work is intended to be dispatched as two documents for publication. Chapters 1 and 2 will be combined to form a research paper, and Chapter 3 is intended to provide the basis for a brief communication. Chapter 1 contains a brief review on ideas about the origin of the genetic code, with a focus on computational analyses. Chapter 2 complements Chapter 1 by presenting commentary and results from analyses on aspects of the genetic code and computational analysis that were unappreciated or neglected in previous studies. Chapter 3 involves testing how well two theories on the origin of the genetic code explain buffering against transcription and translation errors.

Chapter 1

1.1. Background

The origin and evolution of the genetic code remains among the most elusive and contentious mysteries in biology. The genetic code is the rule-set according to which information in DNA (DeoxyriboNuleic Acid) is transcribed to RNA (RiboNucleic Acid), which then is translated into amino acid chains (*i.e*.., polypeptides). In organisms, information is stored in DNA. When organisms, or their constituent cells, utilize stored information, it is transcribed from DNA into complementary messenger ribonucleic acid (mRNA) strands. In ribosomes, mRNA recruits transfer ribonucleic acid (tRNA) molecules, each attached to a specific amino acid. Amino acids are linked together with phosphodiester bonds, forming polypeptides (*e.g*., proteins), which are the functional units in cells.

Nucleotides are arranged into triplets called codons. Each codon is associated with an amino acid (**Figure 1**), but multiple codons may encode the same amino acid. This redundancy is attributable partially to biochemistry (Crick, 1966). The first position in the tRNA anticodon (a triplet that is the reverse complement for its paired mRNA codon) is spatially unconstrained sufficiently to engage in nonstandard base pairing. For instance, the general anticodon GNN can represent the general codons NNC and NNU (wherein N represents any nucleotide). Because these general codons bind (or charge) the same tRNA molecule, they encode the same amino acid. This explanation for redundancy, involving flexibility in the third position, is known as the wobble hypothesis.

The wobble hypothesis can explain only some redundancy in the genetic code. For instance, the wobble hypothesis cannot explain why UUN encodes two amino acids while UCN encodes only one. Moreover, the wobble hypothesis cannot explain why a particular amino acid is assigned to its respective codon.

Complicating the elusive and contentious mystery is the near universality associated with the genetic code. The standard genetic code (SGC; referred to herein as represented – in canonical form – in **Figure 1)** is shared by almost all organisms. Exceptions exist – mitochondria in some organisms (Barrel *et al.*, 1979) and cells in some organisms (Knight *et al.*, 2001) use different codes (**Figure 2**). How the SGC originated and evolved and why it appears to have been altered occasionally remain unsolved.

Another fact beseeching explanation is the observation that similar codons encode similar amino acids (Epstein, 1966). By examining all possible mutations at all positions in all codons, one finds that 24% are synonymous, 39% are nonsynonymous but result in similar amino acids (*i.e.*, amino acids with similar properties; *e.g.*, hydrophobicity or polarity), 8% are nonsynonymous and result in a range in property changes, 22% are nonsynonymous and result in dissimilar amino acids and 7% result in stop (or termination) codons. Most point mutations thus result in negligible effects on protein function.

1.1.1. Frozen Accident Theory (Crick 1968)

Several ideas have been proposed to explain the origin of the genetic code. Among the first major ideas was the frozen accident theory (Crick, 1968). Crick (1968) proposed that a primordial genetic code (PGC) preceded the SGC. The PGC was characterized by some aspects that resembled the SGC. For example, the PGC likely involved triplets. The continuity principle suggests that, if organisms originally used a code with smaller or larger codon sizes, then change to a triplet-based system would have caused all previously existing proteins to have incurred missense or nonsense mutations. This almost certainly would have been lethal, suggesting that such a transition is unlikely to have occurred.

The PGC probably differed from the SGC in other aspects. For example, the PGC might have involved only two nucleotides rather than four and, therefore, encoded fewer codons (Crick, 1968). A reduced codon number and constraints imposed by the wobble hypothesis entail that the PGC would have been unable to have encoded enough proteins to generate functional proteomes. Another aspect, one fundamental to genetic code evolution, is lacking specificity. The PGC likely lacked tRNA molecules that delivered specifically one amino acid for one codon. These primordial tRNAs would have mapped to multiple codons, forming codon blocks, and these codon blocks would have been larger than codon blocks in the SGC. Each codon would have encoded entire amino acid families. Through mutation and natural selection, tRNA molecules gradually would have become more refined.

Effective function by a protein is related to its folded shape, which is determined by properties associated with its constituent amino acids. If a specific codon were to have encoded a variety of amino acids, then some would have resulted in shapes that would have been more effective than would have others; consequently, organisms with mutant tRNAs that consistently recruited more-effective amino acids would have been characterized by fitness advantages relative to their competitors, on a per protein basis. Effects become more complicated in multiprotein systems, wherein changes might have led to some proteins having functioned more effectively and others less; whether such changes would have been selected positively would have determined whether mutant tRNAs ultimately became fixed.

Such a scenario could explain the observation that similar codons encode for similar amino acids. Crick (1968) suggested that the PGC involved a reduced amino acid set and new amino acids were introduced gradually; proteomes must have been small for this to have occurred because replacing amino acids in large proteomes would have imparted devastatingly negative effects on fitness. Crick also suggested that amino acid residues would have been replaced with similar amino acids to minimize disruption. Because new amino acids would have been assigned to codons within larger PGC codon blocks, they would have been surrounded by existing amino acids with similar physicochemical properties. Each new amino acid likely would have been recruited by modifying an existing tRNA. Because tRNAs would have been modified only slightly, cognate amino acids likely would have been related closely to their precursors.

After proteomes had become more complex, the prevailing genetic code would have become "frozen" (Crick 1968). Additional changes to that prevailing genetic code would have been extremely deleterious, most often lethal, so the genetic code no longer would have been able to change, establishing the SGC.

The frozen accident theory provides an explanation for similar codons encoding similar amino acids in the SGC. It also provides an explanation for the SGC appearing to be universal. Since its publication (Crick 1968) and several nonstandard genetic codes having been discovered (Barrel *et al.*, 1979; Maeshiro & Kimura, 1998; Knight *et al.*, 2001) the frozen state description attributed to the SGC has been debunked. The frozen accident theory nevertheless remains influential, and some of its details – such as a PGC evolving into the SGC – are recognizable in the adaptive hypothesis (Sonneborn, 1965) as well as the coevolution (Wong, 1975) theory (described subsequently).

1.1.2. Stereochemical Theory (Woese *et al*. 1966)

When the frozen accident theory was proposed, the major alternative explanation for the origin of the genetic code was the stereochemical theory. The stereochemical theory suggests that each amino acid is associated with a specific codon purely due to stereochemical compatibility. The earliest version was formulated by Woese *et al.* (1966), who observed that some amino acid properties associate with specific nucleotides at specific positions: nonpolar or polar amino acids are associated respectively with codons with a purine or pyrimidine at the second position.

Attempts to explain such associations constitute a long and varied history (Gamow, 1954; Pelc, 1965; Woese *et al.*, 1966; Pelc & Welton, 1966; Melcher, 1974; Balasubramanian *et al.*, 1980; Hendry *et al.*, 1981; Shimizu, 1982). Explanations often involved identifying some physical complementarities between amino acids and anticodons. These attempts have failed to attain prominence in the genetic code literature.

Modern variants on the stereochemical theory have been influenced by the observation that some amino acids bind to RNA (Yarus, 1988, 1991, 1993, 1998, 2000; Yarus & Christian, 1989; Majerfeld & Yarus, 1994, 1998; Majerfeld *et al.*, 2005). The initial breakthrough involved documenting that L-arginine can bind to specific RNA sites (Yarus, 1998). Several binding sites or potential binding sites were discovered subsequently for other amino acids, including L-valine (Majerfield & Yarus, 1994), isoleucine (Majerfield & Yarrus, 1998), tyrosine (Mannironi *et al.*, 2000), tryptophan (Majerfeld & Yarus, 2005), histidine (Majerfeld *et al*., 2005) and glutamine (unpublished work referenced by Yarus *et al*., 2009). Documenting RNA binding sites for the remaining 13 amino acids likely will be achieved in the future. This prediction is supported by tests that show a statistically significant increase in frequencies for codon sequences in binding regions for amino acids (Yarus, 2000), suggesting that codon sequences are important for amino acid binding.

The foregoing observations on RNA-amino acid binding led to the directed template hypothesis. The directed template hypothesis suggests that RNA and amino acids are associated with one another through direct chemical interactions. The directed template hypothesis can take two forms: weak and strong. The weak form suggests that the directed template hypothesis explains only the early PGC, whereas the strong form suggests that the directed template hypothesis explains the SGC. The weak form is complementary with other ideas, specifically the frozen accident theory (Crick 1968), adaptive hypothesis (Sonneborn, 1965) and coevolution theory (Wong, 1975), which propose a PGC wherein codon-amino acid associations existed before the PGC evolved into the SGC.

The weak directed template hypothesis might explain the fact that similar codons encode similar amino acids. If a specific codon were associated with a specific amino acid due to a stereochemical complementarity or chemical attraction, then a similar-but-different codon would be expected to be associated with a similar-but-different amino acid. Such associations could have resulted in the patterns observed in the SGC. While results consistent with the directed template hypothesis are promising, they provide insufficient evidence for supporting the strong directed template hypothesis. Consequently, while the stereochemical theory entails interesting implications for the PGC, it has insufficient support to constitute a compelling explanation for the origin and evolution of the SGC.

1.1.3. Adaptive Hypothesis (Sonneborn, 1965)

Another explanation for the origin and evolution of the SGC is the adaptive hypothesis (Sonneborn, 1965). The adaptive hypothesis suggests that the SGC has been shaped by natural selection to minimize effects from transcription (*i.e.*, mutation) or translation errors. Through degeneracy and codon arrangement, assignments in the SGC ensure that such errors will return either the same or similar amino acids. This ensures that effects on resulting proteins are minimized.

The adaptive hypothesis was first suggested by Sonneborn (1965), who analyzed degeneracy in the SGC. Although the SGC had been deciphered incompletely at the time, Sonneborn identified redundancy in it. He noted that 88% among redundant codons are connected to one another through only 1 nucleotide change. He also noted that some amino acids cannot be connected through any one change. He suggested that this pattern may have been generated by natural selection operating to protect against mutation effects, as such a process would have entailed that mutations resulted typically in either synonymous amino acids or amino acids with similar properties.

Goldberg and Wittes (1966) similarly suggested that the SGC is arranged to minimize deleterious effects from mutation through degeneracy and codon block arrangement. The authors observed that the SGC is buffered more effectively against effects from transition than from transversion mutations and suggested that amino acids requiring more protection (*e.g.*, from thermal degradation) would be expected to contain codons with greater GC content, as they are more stable.

Woese (1965) suggested that the SGC minimizes effects from translation errors (*i.e.*, recruiting incorrect tRNAs). Woese noted that degeneracy is greatest at third positions and errors associated with first positions often lead to substitutions with similar amino acids. Transversion mutations at second positions lead to the most-dramatic translational consequences. Woese remarked that degeneracy levels correspond quantitatively to error rates at each position (10:1:100 for first, second and third positions, respectively; Grunberg-Manago, unpublished results).

Woese *et al.* (1966) implemented a more-qualitative approach to test a claim by Epstein (1966) that some amino acids are more likely to be substituted by other, specific amino acids. By observing distances traveled by amino acids on chromatography paper in a nonpolar pyradine solvent, Woese *et al.* were able to assign a polarity value to each amino acid. They discovered that nonpolar amino acids are characterized by greater codon sequence similarity with other, nonpolar amino acids (and similar patterns characterize polar amino acids). This suggests that the SGC is characterized by an overall 'smoothness' with respect to polarity.

These finding suggest the structure of the SGC may have resulted from non-mutually excluding process(es): physical, chemical and evolutionary. Additional research, now possible at a deeper level than with the foregoing, initial observations, is required to better understand SGC structure.

1.2. Computational Studies on the Genetic Code

One way to evaluate the extent to which the SGC is buffered against transcription or translation error is to compare it against other, hypothetical genetic codes (HGCs). HGCs can be produced by shuffling amino acids identities among codon blocks. This process is accomplished efficiently by computers.

1.2.1. Alff-Steinberg 1969

The earliest such analysis was conducted by Alff-Steinberger (1969). Alff-Steinberger maintained the codon-block structure for the 20 amino acids and three stop codons and compared the SGC to each among 200 HGCs by calculating an error transmission index for each codon position. The error transmission index was the summed absolute pairwise difference over all possible codon mutations for an amino acid property (*e.g.*, GGG-GGA, GGG-GGC, GGG-GGU, … for corresponding amino acid molecular weights). Smaler values indicate genetic codes that are buffered more effectively against errors, whereas higher values indicate genetic codes that are buffered less effectively. Alff-Steinberger examined independently six amino acid properties (molecular weight, polar requirement, number of dissociating groups, PK', isoelectric point and α-helix forming ability) and found that the SGC returned a smaller error transmission index than did the HGCs. Without exception, changes at third positions returned smaller error transmission indices than did changes at first positions, which returned smaller error transmission indices than did changes at second positions. This result accorded with observations that third positions are buffered more effectively against translation errors than are first positions and both are buffered more effectively than are second positions (Woese, 1965). It also supported the notion that the SGC might have evolved in response to relative error frequencies at each position (third>first>second).

1.2.2. Haig & Hurst 1991

Haig & Hurst (1991) used a method similar to that used by Alff-Steinberger (1969). They generated 10000 HGCs through amino acid identity rearrangement (**Figure 3**). Like Alff-Steinberger, Haig and Hurst analyzed each codon position independently and similarly found that errors at third positions were more conservative than were errors at first positions, which were more conservative than were errors at second positions. Unlike Alff-Steinberger, Haig and Hurst used as their error measure mean squared distance. Haig and Hurst examined four properties: polar requirement, isoelectric point, molecular volume and hydropathy. They found that polar requirement and hydropathy were conserved strongly (polar requirement was conserved more strongly), whereas isoelectric point and molecular volume were unconserved. Only one among the 10000 HGCs returned a lower polar requirement score than did the SGC.

1.2.3. Goldman 1993

Goldman (1993) criticized that result by observing that one million represented a tiny fraction among the $2.4x10^{18}$ possible HGCs. This observation complemented observations made previously by Wong (1980) and, to a lesser extent, Di Giulo (1989), who had argued that the SGC was far from the most-effective error-minimizing code. Goldman investigated robustness in the SGC by comparing it to HGCs identified through an efficient, heuristic, error-minimizing, optimization search process. Goldman found that optimized HGCs almost always returned lower mean square errors than did the SGC. Goldman concluded that the SGC was far from optimal and could be improved easily in error-minimization terms. These results later were corroborated by Di Giulo *et al.* (1994) and Judson & Haydon (1999).

Goldman (1993) also conducted a variation on HGC generation by varying codon block structure. Rather than rearranging identities within 20 predetermined codon blocks, Goldman started with 61 blocks (*i.e.*, keeping the three stop codon positions invariant) and allowed each amino acid identity to occupy any available block(s) subject to the constraint that each amino acid identity ultimately had to attain the same codon number as it did in the SGC. Goldman found that no HGC was superior to the SGC under these conditions, ultimately suggesting that codon block structure is an important factor to consider in error measure analyses.

1.2.4. Freeland & Hurst 1998a

Freeland & Hurst (1998a) later would complement the analysis by Haig & Hurst (1991) by accounting for transition and transversion mutation biases (Kumar, 1996) and by accounting for different errors rates in the three codon positions. Despite the potential for twice as many transversions (*i.e*., purine to two pyrimidines or pyrimidines to two purines) as transitions (purine to purine or pyrimidine to pyrimidine), most genomes are characterized by having undergone more transitions. Similarly, errors in the third position of a codon are more likely than are errors in the first position which, in turn, are more likely than are errors in the second position (Woese, 1965). Freeland and Hurst accounted for these biases by weighting more frequent errors greater than less frequent errors . Freeland and Hurst showed in a famously titled paper that only 1 in a million HGCs returned a lower mean square error for polar requirement than did the SGC.

1.2.5 Freeland and Hurst 1998b

In a complementary study to the polar requirement analysis by Freeland & Hurst (1998a), Freeland & Hurst (1998b) tested whether the SGC remained optimized relative to HGCs generated when amino acid shuffling was restricted to members in biosynthetic pathways. The authors ran simulations wherein transitions and transversions were weighted equally and transitions were weighted greater than were transversions. When transitions and transversions were weighted equally, the restricted amino acid set imparted a small but noticeable effect on SGC ranking, approximately doubling it (284 vs. 114 per 1000000). Despite the increase, the SGC ranking remained too low to be considered as a product from chance processes. When accounting for transition and transversion biases, no significant difference in SGC ranking was achieved. Both simulation results suggest that coevolution (*i.e*., between biosynthetic pathways and amino acid incorporation into the SGC) is insufficient to explain the observed error buffering in the SGC.

1.2.6 Ardell 1998

Ardell (1998) used a protein substitution matrix rather than error measure calculations to analyze polar requirement. A protein substitution matrix quantifies how frequently one protein is substituted for another. This can be used as an approximation for amino acid similarity because more-similar amino acids will be substituted interchangeably at a higher rate relative to less-similar amino acids, as substitutions among more-similar amino acids impart smaller impacts on protein function. Ardell primarily was concerned with testing whether the factors that effected buffered codes could have operated on specific positions within codons. Ardell found that position-invariant factors, like base-content and mutation, could explain code optimization and thus error-minimization might have been involved only early in SGC evolution.

1.2.7. Judson & Haydon 1999

Judson & Haydon (1999) evaluated how effective the SGC was relative to HGCs produced in simulated evolutionary processes. They created a computer program involving a genetic algorithm that generated HGCs, allowed them to mutate and 'breed' (*i.e.*, create offspring codes with properties from both parental codes) and selected the HGC with lowest error measure score in each population to continue a lineage. Error measure scores were calculated as absolute and squared differences for eight physicochemical properties and six derived variables for eight structural properties. Over several generations, the most-refined HGC was compared with the SGC. Judson and Haydon found that HGCs produced in this manner consistently outperformed the SGC at all three codon positions. The authors used their genetic algorithm to select for HGCs with similar structure to that in the SGC. Structure was assessed on the basis of connectedness: a

HGC was found to be similar to the SGC if both would yield similar synonymous, nearly synonymous (*i.e*., leading to similar amino acids) and non synonymous mutation numbers. Under these conditions, most-refined HGCs were much more similar to the SGC. Judson and Haydon suggested that the SGC ultimately resulted from selection operating on a primordial code with flexibility and potential for adaptation.

1.2.8 Freeland et al. 2000

Freeland *et al.* (2000) disagreed with the suggestion that identifying more-effectively buffered HGCs is tantamount to rejecting the hypothesis that the SGC resulted from an evolutionary process in which natural selection operated to minimize error. They considered two situations: one where the SGC was allowed potentially to compete against the $2.4x10^{18}$ possible HGCs and one where competition was restricted to a smaller ($\sim 10^9$ possible) HGC subset delimited by biosynthetic associations. They also used data from protein substitution matrices to quantify similarities among amino acids (and compared these to analyses involving polar requirement to relate their results to previously published results). They found that the SGC was between 76% and 97% optimized relative to the unrestricted set and between 96% to 100% optimized relative to the restricted set. Freeland *et al.* conceded that imagining a real-world scenario wherein optimal alternative genetic codes could have evolved and competed, ultimately producing the SGC was challenging because, as increasingly optimized codes occur infrequently, competition among such codes is unlikely to have occurred. As error-minimization became more enhanced, evolutionary rates would have decreased, and codes would have become frozen before a global error minimum could have been achieved.

1.2.9 Ardell & Sella 2001

Ardell & Sella (2001) investigated why the SGC is restricted to 20 amino acids. Ardell and Sella conducted computer simulations that began with an ambiguous code, which gradually evolved into a more-specified code through successive alterations. Fitness for codes was determined by examining the ability to translate a message. Each message involved sites, and each amino acid was assigned a fitness value for each site. Amino acids, here, referred to 20 hypothetical amino acids created for the simulation. Each amino acid was assigned a value ranging from 1 to 20, which served as a proxy for its physiochemical properties. Owing to their different values, some amino acids were suited better at a particular site than were others. The more dissimilar an amino acid value was from the optimal amino acid value at a particular site, the greater was the fitness cost for the dissimilar amino acid occurring at that site. The actual fitness cost was determined by the user as well as the probability for mutation (*i.e.*, to a new identity). Each mutated code fitness would be compared to the nonmutated code by its ability to faithfully translate a message. New mutated codes were allowed to compete with predecessor codes, and most-fit codes outcompeted less-fit codes. The simulations mimicked a natural selection scenario for how the SGC might have evolved. They found that codes always became nonambiguous and almost all codes became redundant, as observed with the SGC. The final encoded amino acid number resulted from a balance between mutation rates and toleration for missense mutations. A higher mutation rate encouraged more redundancy whereas a higher missense toleration produced greater diversity. The codon number for the SGC may have resulted from that interplay. The authors moreover found that the amino acids incorporated into evolving codes were characterized by moderate physiochemical characteristics. These amino acids may have been more fit because they could function in a wide variety of roles.

1.2.10 Gilis *et al.* 2001

Gilis *et al.* (2001) investigated the effects that amino acid frequencies impart to error buffering in the SGC. Amino acids occur heterogeneously among proteins; mutations involving more-frequent amino acids impart a greater effect on fitness because they occur more frequently. Amino acid distances were determined two ways: PAM matrices and a cost matrix. The cost matrix estimated the free energy change in a protein after substituting one amino acid for another. The estimation was obtained by measuring torsion potentials with non-localized long range hydrophobic interaction data. The authors found that the SGC was even more optimized than previously thought. Among 10^9 HGCs, the SGC ranked second.

1.2.11 Goodarzi *et al*. 2004

Goodarzi *et al*. (2004) considered possible effects from stop codons. Mutations to stop codons, nonsense mutations, truncate proteins, which almost always nullify function. Goodarzi *et al.* examined error buffering in the SGC while accounting for effects from stop codons and using a fitness measure derived from amino acid substitution matrices. The authors ran a variety of simulations in which different penalties for mutating to a stop codon were assigned. Rather than calculating distances, nonsense mutations instantly incurred penalties. Goodarzi *et al.* also incorporated amino acid frequencies into their calculations. Mutations resulting from more-frequent amino acids were weighted higher than were mutations from less frequent amino acids. They discovered that the SGC returned the lowest error measure score for every 10^9 HGCs examined. As with other investigations, stop codon positions notably remained frozen in place.

1.2.12 Vetsigian et al. 2006

An explanation related closely to the hypothesis that the SGC resulted from an evolutionary process in which natural selection operated to minimize error was presented by Vetsigian *et al.* (2006), who suggested that optimization occurred through horizontal gene transfer in pre-LUCA (Last Universal Common Ancestor) communities (Woese *et al*., 1990). These communities would have contained multiple unicellular individuals with distinct genetic codes. These individuals frequently would have experienced horizontal gene transfer. The authors suggested that, for individuals to have utilized effectively donor-protein innovation, recipient genetic codes would have to have resembled closely donor genetic codes. The closer the resemblance, the more-effective would have been the innovation. Consequently, whenever organisms received material, natural selection would have operated, favouring organisms with recipient genomes that were similar to donor genomes. Gradually, this would have drawn together all codes into a single, universal code. Vetsigian *et al.* tested this scenario through computer simulation. By using a process similar to that used by Ardell and Sella (2001), Vetsigian *et al*. allowed code populations to evolve over time. Like Ardell & Sella, Vetsigian *et al.* measured the fitness for a code by constructing a hypothetical message where codons were assigned to certain site type. Each site type had an optimal amino acid, with other amino acid resulting in a reduction in fitness proportional to the actual amino acid and the ideal amino acid. Unlike Ardell & Sella (2001), the amino acids used were real amino acids whose similarity was determined through Hamming Distances (amino acid differences between two proteins). As simulations progressed, mean distance between neighboring amino acids as well overall similarity among codes in the population were evaluated without and with horizontal gene transfer. The authors found that horizontal gene transfer resulted in genetic codes becoming more similar over time, in contrast to outcomes in populations evolving without horizontal gene transfer. Although populations under both scenarios tended toward reduced mean distances between neighboring amino acid properties, the process was more effective in populations evolving with horizontal gene transfer. This suggests a plausible alternative to an already universal code become refined through natural selection.

1.2.13 Novozhilov *et al*. 2007

Novozhilov *et al.* (2007) also investigated how well buffered against error the SGC is when compared to possible HGCs within a restricted space among all possible HGCs (containing genetic codes with corresponding block structure and degeneracy to the SGC). Novozhilov and colleagues used a cost function that measured the frequency of that amino acid being considered, the cost of substituting two amino acids and the position of the codon being altered. As did Goldman (1993), they found that the SGC was far from optimal. They generated pseudorandom codes and conducted simple, pairwise exchanges between amino acid identities in two- and four-codon blocks, ultimately to generate fitness increases. On average, the SGC was buffered more effectively than were the HGCs and required less time to reach fitness maxima. However, when starting with HGCs already characterized by error measure levels that were similar to the SGC, refined HGCs with greater resistance to transcription and translation errors could be evolved easily. These results suggest that the SGCs experienced partial optimization before becoming frozen.

1.2.14 Tlusty 2010

Tlusty (2010) also attempted to investigate why the SGC assigns similar amino acids to similar codons and why it contains 20 amino acids. Tlusty considered three separate factors involved in the origin and evolution of the SGC: error-tolerance, diversity and costs associated with biochemical machinery. Error tolerance was measured by looking at the sum over all possible substitutions, accounting for substitution frequency and distances between amino acids. Distance was measured on the properties polar requirement and molecular volume. Tlusty suggested that, when the SGC was shaped by natural selection, fitness was maximized over these three factors. One illuminating contribution by Tlusty involved using the map coloring problem. The map coloring problem refers to the coloring-number required to identify areas in a map so that no two adjacent areas have the same color; with respect to genetic codes, the coloring problem can be utilized to determine how many amino acids can be encoded while maintaining a 'smooth' (*i.e.*, low error measure variation) topology. The maximum coloring number was found to be 25 if all 64 codons are free to vary and 20 if restrictions imposed by the wobble hypothesis are included. This finding may explain why the SGC contains only 20 amino acids.

1.3.1 Implications for Ideas on Genetic Code Origins

The foregoing review shows that, depending on how one measures and interprets robustness, computational analyses might be interpreted as indicating that the SGC resulted from either an astounding optimization process in which the SGC was buffered against transition or translation errors or modest improvements on arbitrary genetic codes that may have been by-products from other factors. Given the vast codon space available and the fact that error-minimization becomes increasingly difficult to improve as HGCs become more and more buffered, a compromise perspective probably is most likely, and the SGC is considered most appropriate as partially optimized, consistent with the adaptive hypothesis (Novozhilov *et al.*, 2007). The SGC might be improved easily in error minimization terms, but the main implication from error measure analyses is that it is buffered more than would be expected by chance, and this pattern warrants explanation.

1.3.2 Unfreezing the Genetic Code

The adaptive (*i.e*., error-minimization) hypothesis has been criticized. Crick (1968) suggested that, once a genetic code had been established, any additional changes would have been lethal. Although variant genetic codes might have been more-buffered against transition and translation error, such genetic codes would have been at a severe disadvantage. Variant codes would have encoded different amino acid compositions for existing, functional proteins. These variant proteins almost certainly would have functioned less effectively. Variant genetic codes consequently would have been eliminated quickly by natural selection.

In response to this suggestion, several researchers have proposed alternative mechanisms for how the SGC originated and evolved. One mechanism (Osawa & Jukes, 1989) involves the notion that codons changed their amino acid identity assignments, a process called codon swapping. Essentially, mutational biases would have caused some codons to fall into disuse. These codons could have mutated neutrally to encode new amino acids. Szathmary (1991) showed that this was a feasible mechanism for code evolution. Another mechanisms involves the notion that, in the earliest genetic code, codon assignments may have involved fewer amino acids (Trifonov 2000, 2004; Higgs and Pudritz 2007, 2009; Higgs, 2009); one particular version suggests that four amino acid assignments constituted the initial configuration (Higgs 2009), which could be represented in a one-to-one manner as four columns **(Figure 1**). The four columns ultimately would have become subdivided into the codon blocks in the SGC, containing the remaining 16 amino acids and stop codons. New amino acids were incorporated only if the newly encoded assignments minimally disrupted existing proteins and pathways and ultimately conferred a fitness benefit (in direct competition with each other or to the individuals containing them). In this way, the code could have evolved ‘bottom-up.’

1.3.3. Coevolution Theory

Another criticism that could be levelled against the adaptive hypothesis involves other, more-parsimonious explanations for SGC structure. The most-popular explanation was formulated originally by Wong (1975) and is known as coevolution theory. Coevolution theory suggests that the PGC comprised fewer amino acids with greater codon redundancy. These precursor amino acids were modified over time to create the remaining amino acids. For example, aspartic acid currently is modified to produce asparagine in one step, through ATP-powered amination and simultaneous glutamine-to-glutamate transformation (Milman & Cooney, 1979). Several other amino acids are produced through similar biosynthetic pathways (Taylor & Coates, 1989). Amino acids related in this manner are said to hold product-precursor relationships. If coevolution theory is correct, then one would expect product amino acids to be encoded by codons formerly assigned to precursor amino acids; consequently, product-precursor amino acids would be located close to one another in the SGC. Such product-precursor amino acids, in fact, often are in close proximity. Wong calculated the probability for the observed relationships as statistically significant.

No definitive way to group amino acids into biosynthetic families exists (Amirnovin, 1997). Wong (1975) identified several product-precursor pairs (**Table 1**), and these can be organized into biosynthetic families. Di Giulio & Medugno (2000) subsequently identified five biosynthetic families (**Table 2**), with a slightly different composition than the patterns that Wong originally identified. One would expect that product-precursor amino acids would have similar properties, and so the SGC would comprise codon blocks containing similar neighboring codons. Wong (1975) suggested that multiple codes wherein product amino acids were assigned to different codon sets derived from precusor amino acids might have existed. These codes could have competed against one another, and the code that best protected against transition or translation error (and, consequently, which bore the greatest fitness) would have become fixed.

Wong (1981) proposed a scenario wherein amino acids became incorporated in three phases. Phase I amino acids were available via prebiotic synthesis and could have been incorporated into cells immediately. Phase II and III amino acids were constructed through inventive biosynthesis and post-translational modification, respectively. Phase II and III amino acids would have been unavailable for incorporation into the PGC immediately, and, so, they would have had to have been added to the PGC later. Wong notes that error-minimization might have been involved in SGC formation but claims that it played a minor role, subsidiary to the processes involved in coevolution theory.

1.3.4. Biosynthetic Pathways

While coevolution theory remains a compelling explanation for SGC structure, its statistical basis was criticized by Amirnovin (1997). Amirnovin suggested that most amino acids are related biosynthetically to one another, and, so, Wong (1981) finding statistical significance for biosynthetically related amino acids inhabiting adjacent positions in the SGC codon table is unsurprising. Amirnovin suggests, then, that claiming that SGC structure supports coevolution theory is premature. He decided to demonstrate this computationally by generating HGCs and exploring how closely related they were. He found that similarity was highly dependent on which amino acid pairs were considered to be related biosynthetically. Depending on the amino acids considered, the probability for meeting or exceeding the relatedness level observed with the SGC ranges from $p=0.001$ to $p=0.34$.

These results were contested by Di Giulio and Medugno (2000), who found fault with the methodology. They argued that quantifying the relationship between two amino acids using one-parameter significance was flawed. Such a method inadequately captures genetic code structure. Di Giulio and Medugno argued that a hypergeometric distribution should be used instead. They also took exception to the way that Amirnovin calculated the probability for two biosynthetically related amino acids to be adjacent to one another in codon space. This value, the codon correlation score (CCS), was quantified by counting how many ways one amino acid could mutate into another amino acid through point mutation. Arminovin generated a HGC set and used it to determine the probability for achieving a certain relatedness score. Di Giulio and Medugno suggested that a more-appropriate measure would involve calculating the "probability of observing a certain CCS value on the condition that the CCS value is produced only by those [codes] that have a number of amino acid pairs at least equal to that of the pairs that are effectively specified in the genetic code and whose significance has to be established." The authors corrected for this by deciding "to exclude the random codes that have the same rare CCS value but which are actively produced only by a number of amino acid pairs lower than the number of pairs effectively specified in the genetic code and whose significance has to be established". After addressing these issues, they calculated the probability for meeting or exceeding the relatedness level observed in the SGC as $p=10^{-6}$.

This rebuttal generated a surrebuttal by Ronneberg *et al.* (2000), who argued that previous methods for calculating similarity also were flawed. Ronneberg *et al.* argued specifically that previous authors had used incorrect product-precursor relationships and failed to account for restrictions imposed by processes associated with the wobble hypothesis. After addressing these issues, the authors calculated that probability that the observed biosynethetic relationships in the SGC were achieved due to chance was p=0.62.

One possibility, originally alluded to by Wong (1975), is that coevolution and error-minimization occurred together, and, moreover, the initial PGC may have been determined at least partially by stereochemical affinity. Wong (2005) estimated that the relative contribution by biosynthetic pathways, error-minimization and strereochemical affinity in SGC evolution respectively was 40000000:400:1. Di Giuilo (2005) suggested that precursor codons conceded codons to product amino acids such that product amino acids arranged themselves in columns. Another possibility is that codons were repositioned after their initial assignment by codon swapping to result in a more-buffered code.

The idea that coevolution and error-minimization may have played important roles in SGC evolution has been supported by statistical analyses. Szathmary & Zintzaras (1992) compared similarities for tRNA molecules between biochemically related amino acids and physicochemically similar amino acids. They suggested that, if tRNAs were more similar when correlated biochemically, then coevolution predominated; whereas if tRNAs were more similar when correlated phyiscochemically, then error-minimization predominated. They found a greater correlation among tRNA molecules was achieved on the basis of physicochemical similarity. The authors concluded that, while both processes played important roles in SGC evolution, error-minimization played a larger role in organizing the SGC. In addition to initial codon assignments – potentially guided by coevolution – codon reassignments contributed to SGC structure.

Chapter 2

2.1. AGCT

The tool that we developed and used to perform genetic code error measure analyses is a computer program called AGCT, written using the technical computing environment *Mathematica* (Wolfram Research, Inc. 2010; v. 9.0.1.0) as a software platform. AGCT conducts analyses in a manner similar to the approach adopted by Alff-Steinberger (1969), Haig & Hurst

(1991), Freeland & Hurst (1998a, 1998b), Freeland *et al.* (2000), Gilis *et al.* (2001) and Goodarzi *et al.* (2004).

AGCT generates hypothetical genetic codes (HGCs) by shuffling amino acid identities among codon blocks and calculates change associated with every possible point mutation at every possible codon position as a distance. Each codon potentially is subject to nine point mutations, although users may opt to exclude those that lead to stop codons. Distances are evaluated for each codon, and the average value over all codons is determined to create a single value for an entire code. Codes that return lower values are buffered against transcription and translation errors. To discover where the standard genetic code (SGC) resides among the HGCs, AGCT sorts all distance values and identifies where in their distribution the distance for the SGC resides.

2.1.1. Distance Metric Calculations

Each amino acid is characterized by quantifiable properties, such as hydrophobicity and polarity. The property value distance between two amino acids associated with codons in the SGC and the property value distance for the two amino acids associated with the same codons in a HGC can be compared. A larger difference indicates that the amino acids associated with the codons in the HGC are more dissimilar. One cannot merely compare genetic codes by taking differences between all possible values, however. The natural symmetry over all pairwise comparisons would lead to a null total for every genetic code. AGCT provides users two options to circumvent this symmetry: calculating absolute (*e.g.*, Alff-Steinberger, 1969) or squared (*e.g.*, Haig & Hurst, 1991) distances, respectively AD or SD.

2.1.2. Stop Codons

AGCT notably is the second computational tool that allows stop codons to be included in error measure calculations (*i.e.*, in addition to the one used in Goodarzi *et al.*, 2004). When stop codons are included in calculations, AGCT treats each as it does amino acids. By default, the value assigned to each stop codon for each property is 0. This value is arbitrary but could, and typically does, impart relatively great effects depending on the property under consideration (*i.e.*, how values among amino acids are distributed) and HGC structure (*e.g.*, whether substantially more stop codons are assigned to triplets relative to the SGC). A natural complement to a 0 value would be to determine minimum and maximum possible errors (*i.e.*, among all possible pairwise amino acid comparisons) for any property under analysis and assign those values to all calculations involving stop codons for that property, to return minimum and maximum distances. Analyses then could be performed using those minimum and maximum distances, which would provide a bracketing interval for stop codon effects in error measure calculations.

AGCT is the first computational tool that allows stop codons to be included in codon-identity shuffling. Similar to the two amino acid identities that occupy noncontiguous codon blocks in the SGC but are considered as singulars during shuffling (*e.g.*, serine and arginine), the two vertically contiguous stop codon block positions (*i.e.*, UAA and UAG) are considered as a singular entity.

2.1.3. Properties

AGCT was developed with the principle aim to expand on analyses already conducted. Previous investigations were limited in two main ways: they utilized only a limited variety of amino acid properties and they considered each property only in isolation. For instance, Alff-Steinberger (1969) analyzed the SGC with respect to 6 properties independently (*i.e.*, molecular weight, polar requirement, number of dissociating groups, pK', isoelectric point and a-helix forming ability). Subsequent investigations (Haig & Hurst, 1991; Goldman, 1993; Ardell, 1998; Freeland & Hurst, 1998a, Freeland *et al.*, 2000; Goodarzi *et al.*, 2004; Higgs, 2009; summarized in **Table 3**) either focused on these properties with a few additions or solely focused on the polar requirement, which is related to hydrophobicity (Woese *et al.*, 1966).

AGCT is the first computational tool that allows comparisons over multiple properties simultaneously. One criticism that could be levelled against adopting this approach is that the scale for variation in each property essentially determines the average distance magnitude for that property relative to other properties. Comparisons could return results biased by properties with relatively large values or variances. One way to address this criticism is to transform distance values for different properties so that they can be compared on the same scale. AGCT utilizes standard scores (*i.e.*, mean-zero and standard deviation unity) for such comparisons, a transformation that has been used previously to unify multiple properties in a different analytical context (Higgs, 2009). When this option is enabled, ACGT reports the standard score for the SGC or any HGC relative to the entire standard score distribution rather than the mean AD or SD score for each property. The standard score ranking for a genetic code indicates how effectively that genetic code minimizes errors for a particular property relative to the entire distribution. A genetic code that is buffered effectively for a particular property should return an extreme negative value, whereas one that is buffered less effectively should return a value closer to zero or even extending into positive values. Standard scores over all properties are added together to obtain a unified score for a genetic code. Genetic codes that are buffered effectively over multiple properties should return extreme negative unified scores, whereas those ineffectively buffered should return large positive unified scores. Genetic codes that are buffered effectively for some properties while ineffectively for others should return unified scores closer to zero.

Another criticism that could be levelled against adopting the approach encoded in AGCT and error measure approaches generally involves HGC space. Many HGCs can be generated by shuffling codon block identities: $2.43*10^{18}$ and $2.59*10^{22}$, when excluding and including stop codons, respectively. These numbers dwarf the HGC set sizes that typically are used in studies. A particular HGC set therefore fails to represent the entire theoretically possible distribution. This problem has been considered previously. Freeland and Hurst (1998a) compared results obtained from 10000 and 1000000 HGC sets. They found that results were stable between the two conditions, suggesting that 10000 HGC sets are sufficient to create a representative sample.

To confirm this finding, we ran 10 replicates with 10000 HGC sets to analyze the property hydrophobicity. This property was chosen because it is conserved modestly, which allowed the SGC to vary in ranking among replicates. As each HGC in each replicate was generated independently using pseudorandom processes, if each 10000 HGC set was nonrepresentative for the entire possible distribution, then the standard deviation for the HGC set would be large; conversely, if each set was representative for the entire possible distribution, then the standard deviation for the HGC set would be small.

The SGC ranking was very stable among replicates (**Figure 4**), with a mean ranking ≈3361 and a standard deviation ≈31. Theoretically, given a corresponding infinite normal distribution, 97% observations reside between values 3299 and 3423. This range represents only 1.24% (124/10000) among the possible values. Consequently, when the SGC ranking varies among replicates, changes are minor and quantitative rather than qualitative in nature. On the basis of this result involving a modestly conserved property, we consequently and cautiously are confident that implementing 10000 replicates provides reasonable representation for entire parameter spaces while allowing efficient exploration over the 54 variables (derived from Grohima *et al.*, 1999 & Higgs, 2009) analyzed herein.

2.2. Analyses

2. 2.1. Distance Metric Calculations

In addition to investigating previously unexamined amino acid properties, we also sought to investigate how different distance metrics affect SGC ranking in error measure analyses. Both distance metrics, AD (Alff-Steinberger, 1969) and SD (Haig & Hurst, 1991), have been used previously to quantify error, but SD has been used more frequently. Both metrics are valid *a priori*, as they address the previously mentioned symmetry issue. We ran 10 replicates twice, each replicate containing 10000 HGCs. We thereby formed two sets, one for analysis with AD and one for analysis with SD. Both sets involved 6 properties (hydrophobicity, polar requirement, isoelectric point, bulkiness, surface area accessible to water when unfolded and fraction of accessible area lost when a protein folds) transformed into a single, unified score.

We found that, contrary to expectation, distance metric affected greatly results obtained (**Figure 7)**. The AD and SD metrics produced two distinct populations. The SGC ranked lower in the AD population than it did in the SD population. The SD metric involves squared differences. Larger differences will impart greater effects on error measure calculations, which might affect SGC ranking. This effect is dependent on HGC code structure. The SGC is arranged with similar amino acids in columns. This means that most transcription or translation errors (*i.e.*, those associated with first or third codon positions) would produce either the same or similar amino acids; some errors (*i.e*., those affecting second positions) would produce different, dissimilar amino acids. A HGC that was characterized by greater differences between vertically neighboring amino acids would return greater error measures than would the SGC, which consequently would ranking low. The SGC, however, is characterized by a few instances where vertically neighboring amino acids are extremely different for some properties. Squaring the associated distances might return an error measure sufficiently large to rank the SGC relatively high among HGCs.

2.2.2. Stop Codons

One computational aspect unique to AGCT is its ability to allow users to consider completely effects imparted by stop codons. Previous investigations on the SGC either ignored stop codons entirely or assigned an error function quantifying mutation to stop codons while keeping stop codon blocks and positions fixed (Goodarzi *et al.*, 2004). We analyzed stop codons with two two-state options: fixing or varying assignment positions and excluding or including mutations involving stop codons in error measure calculations. We performed three analyses, each involving 10 replicates containing 10000 HGCs, with SD as the error measure. We investigated polar requirement, as this property has been analyzed thoroughly by previous researchers (*e.g.*, Alff-Steinberg, 1969; Haig & Hurst, 1991; Szathmary & Zinteras, 1992; Goldman, 1993; Ardell, 1998; Freeland & Hurst, 1998a, 1998b; Judson & Haydon, 1999; Higgs, 2009; Tlusty, 2010):

I. stop codon positions fixed, included in calculations;
II. stop codon positions variable, excluded in calculations;
III. stop codon positions variable, included in calculations

(the fourth possibility – stop codons fixed , excluded in calculations – has been explored exhaustively in previous error-buffering investigations).

Using Welch's t-test, we found significant differences in ranking between conditions I (μ=2.10, SD=1.04) and II(μ=3.55, SD=1.93) (t_{30}=2.93, p=0.006), and no significant differences between conditions I (μ=2.10, SD=1.04) and III (μ=1.70, MD=0.80) (t_{35}=1.34, p=0.19). These findings, together, may be interpreted as indicating that the difference between conditions I and II can be attributed predominantly to including or excluding stop codon in error measure calculations rather than fixing or varying codon positions. This interpretation is supported by our finding significant differences in ranking between conditions II (μ=3.55, SD=1.93) and III (μ=1.70, MD=0.80). When stop codons were excluded from calculations, the SGC ranked higher (*i.e.* was less conserved relative to HGCs) than it did when stop codons were included (t_{25}=3.95, p=0.0005). We hypothesized that the SGC fares better at error-minimization when stop codons are included because most HGCs contain more stop codons than does the SGC, leading to more, potentially disfavorable distances in error measure calculations.

We tested this hypothesis. We ran another analysis, involving 1000 HGCs, with SD as the distance metric, in which AGCT tracked stop codon mutations for each HGC. We again investigated polar requirement. We predicted that error measure scores would increase with increases in stop codon mutation number. We found no relationship between SGC ranking and stop codon mutation number when stop codons were excluded from error measure calculations, as expected (**Figure 8**). When stop codons were included in error measure calculations, however, we observed a significant, positive relationship (**Figure 9**). This suggests that the SGC ranks low in error measure analyses partly because stop codon blocks in the SGC are small and partly because polar requirement values among amino acids are distributed so that the 0 value assigned to changes involving stop codons imparted large effects.

An informative follow-up study would involve weighting mutations to stop codons according to codon assignment frequency. This approach has been conducted partially by Goodarzi *et al*. (2004) albeit with invariant stop codons. If stop codons were surrounded by comparatively rare amino acids, then the associated nonsense mutations would be even less likely to occur. Adding this factor may allow researchers to quantify and analyze stop codon position effects in addition to stop codon frequency effects.

2.2.3 Properties

2.2.3.1. Individual Properties

Previous computational error measure analyses were focused on a narrow range of properties. We therefore considered an expanded list for analysis. We added to four commonly studied properties (hydrophobicity, polar requirement, absolute entropy and melting point) 45 properties described by Grohima *et al*. (1999; listed in **Table 4**) as important for stabilizing folded proteins.

These 49 properties are nonorthogonal and several properties intuitively are expected to be correlated. For example, hydrophobicity would affect strongly buriedness (the more hydrophobic an amino acid, the stronger its tendency to retreat to the interior in a folded protein to minimize its contact with surrounding water molecules; Tanford, 1962)). The properties can be grouped into four major categories: hydrophobicity, enthalpy, Gibbs free energy, and miscellaneous. Properties related to hydrophobicity and enthalpy generally are conserved in error measure analyses, whereas those related to Gibbs free energy generally are unconserved. That the SGC effectively minimizes errors associated with polar requirement has been established convincingly (Haig & Hurst, 1991), so this property serves as a useful reference.

To explore properties individually, we used SD as the distance metric and excluded stop codons in calculations. We report four main observations. First, strongly conserved properties included (in rank order with descriptors provided by Grohima *et al.* (1999) unless otherwise specified): El (long-range non-bonded energy), Hgm (combined surrounding hydrophobicity), -

TΔSc (unfolding enthalpy change of hydration), polar requirement, Nl (long range contacts), Hc (unfolding enthalpy), Rf (chromatographic index; Woese, 1966), ΔASA (solvent accessible surface area for unfolding), Cph (unfolding hydration heat capacity chance), Ns (average number of surrounding amino acid residues when inside a folded protein), Ra (solvent accessible reduction ratio), Br (buriedness), Hp (surrounding hydrophobicity), Et (total nonbonded energy) and F (root mean square fluctuation displacement).

Second, the SGC effectively minimizes errors for properties mediating long-range interactions (El, Nl), whereas it ineffectively minimizes errors for properties mediating short-range interactions (Esm, Nm).

Third, more properties are conserved than expected by chance. If the SGC structure failed to minimize errors associated with transcription or translation, then one would expect that approximately half the HGCs would score better and half would score worse, barring effects from sampling bias. We discovered a strong skew among properties, toward effective buffering. This pattern can be explained by the aforementioned nonorthogonal relationships among properties. Several are affected by polarity, for instance; if polarity is conserved, then all affected properties also are conserved. This, however, cannot explain the observation that properties in different categories are conserved, which is consistent with the idea that SGC structure buffers against transcription and translation errors.

Fourth, long range non-bonded energy (E1) was the most conserved property in our analysis. This is noteworthy because polar requirement often has been reported as the most conserved property (Alff-Steinberger; Haig & Hurst, 1991). We therefore decided to explore this property more thoroughly. We compared the SGC to 100000 HGCs and included stop codons in calculations. We found that the SGC achieved rank 2 (**Figure 5**). For reference, Haig & Hurst (1991) found the SGC achieved rank 1 among 10000 HGCs for polar requirement.

We consider three ways to interpret this result. The first, and most dramatic, is that the SGC became organized to minimize errors associated with long range non-bonded energy rather than some other property (*e.g.*, polar requirement). The extent to which polar requirement errors are minimized, indeed, may result from the nonorthogonal relationship between these two properties. This explanation is unlikely to achieve consensus, as polar requirement is more important in protein folding. A second explanation is that the SGC became organized to minimize errors associated with polar requirement and the extraordinary buffering in long-range non-bonded energy is a byproduct from that constraint. Establishing relative error effects on protein products as well as quantifying their interrelationship would help determine which property should be considered as primary in SGC buffering. A third explanation is that the SGC became organized to minimize errors associated with both properties, as both may be crucial for proper folding in the protein product. This would suggest that the SGC is more buffered than previously suspected.

2.2.3.2. Multiple Properties

The foregoing analysis suggests that, although individual properties like polar requirement are important for protein folding, they independently fail to capture entirely the error buffering capacity for the SGC. Two approaches may be adopted to attain a more-complete perspective: analyzing either parameters that account for many properties together or multiple properties simultaneously.

The first approach has been attempted by implementing point accepted mutation (PAM) matrices (*e.g.*, Ardell, 1998; Freeland *et al.*, 2000; Ardell & Sella, 2001; Gilis *et al.*, 2001;

Goodarzi *et al.*, 2004; Novozhilov *et al.*, 2007). A PAM matrix is constructed with data obtained from homologous proteins and quantifies how often one amino acid has been observed to have substituted for another. As changes between similar amino acids are more likely to be considered neutral (and, therefore, less likely to be selected against), a higher pairwise substitution rate should indicate that two amino acids are more similar to each other than are two amino acids associated with a lower substitution rate. A notable exception might occur when change between dissimilar amino acids would confer a selective advantage. Such instances are expected to occur rarely (Kimura 1968) and, then, within specific families for homologous proteins, at specific positions; consequently, this phenomenon will be blunted when sampled over many sites and substitution rates should reflect accurately similarity between proteins.

Although PAM matrices thus would seem to present an ideal way to analyze a single parameter representing multiple properties, PAM matrices are plagued in studies on the SGC by a critical flaw: values contained in PAM matrices reflect the SGC structure, itself (Freeland et al., 2000; Higgs, 2009). Adjacent amino acid identities differ by single nucleotides, whereas remote amino acids differ by more than a point mutation or misread nucleotide. Amino acids that differ by a nucleotide, between which neutral mutations may occur, should be characterized by greater substitution rates; conversely, amino acids assigned to codons that differ by more than one nucleotide, between which mutations are less likely to occur, should be characterized by smaller substitution rates. PAM matrices and SGC structure thus are nonindependent.

A second approach to performing more-holistic error measure analyses would be to consider multiple properties simultaneously. This approach has multiple disadvantages. Researchers must select which properties to include, whether to weight them and, if so, then by how much. As mentioned previously, many properties are interrelated, so including them effectively weights some more than others. For instance, if, along with polar requirement, five properties were to be included in an analysis, with two correlated strongly with polar requirement and the other three orthogonal to polarity and one another, then polar requirement effectively would have been considered thrice. One also may err by failing to include all the properties needed to capture the essence for error buffering. All these disadvantages are compounded by the fact that protein folding is understood incompletely, which may prevent proper property selection.

Some properties (such as polar requirement) are more important to protein structure and function than are others (such as isoelectric point). One might consider assigning greater weight to more important over less important properties. Deciding on values for weightings is arbitrary and challenging, and, without adequate information, weighing properties may return inaccurate results, as erroneous as failing to have weighted properties at all (or even more).

These factors ensure that, until greater understanding about protein folding has been achieved, error measure analyses involving multiple properties should be conducted cautiously. The approach nevertheless still has merit. Although one might be unable to account effectively for every property, this limitation would be true with any approach. Moreover, given that many properties are interrelated, any unaccounted for property may be represented indirectly by other, included properties. We opted to use multiple properties over a PAM matrix for two reasons: to avoid the aforementioned circularity and to generate results that complement those obtained with analyses using PAM matrices (Ardell, 1998; Freeland *et al.*, 2000; Ardell & Sella, 2001; Gilis *et al.*, 2001; Goodarzi *et al.*, 2004; Novozhilov *et al.*, 2007).

We suspected that the SGC might be more than the sum-of-its-parts in error buffering terms. We note that previous error measure analyses have revealed that some traits are conserved effectively and others less effectively (*e.g.*, polar requirement is conserved effectively; Alff-Steinberger, 1969; Haig & Hurst, 1991; Freeland & Hurst, 1998a, 1998b). This might indicate the SGC is organized to minimize errors for those particular, conserved traits; alternatively, the SGC might be organized to elicit a negative interaction over multiple properties (*i.e.*, one that reduces total error). If the SGC were organized to minimize effects from relatively few, specific, conserved properties, then a unified score including many other properties would rank somewhere between the rankings for the most conserved properties and the rankings for the least conserved properties; alternatively, if the SGC were organized to elicit a negative interaction over multiple properties, then a unified score would rank lower than expected on the basis of compromise – in the most extreme circumstance, less than or equal to the ranking for the most conserved property (with equivalence at rank 1). In this instance, if a HGC were found to rank lower for a given property (*e.g.*, its SD score were 1 unit less than the SGC), then we would expect that this would be counteracted by an increased ranking in the SD score for at least one other property (*e.g.*, its summed SD scores over the remaining properties would exceed 1 unit). Using standard (rather than raw) scores might constrain the effect that negative interactions would impart and, so, their identification.

We generated data to distinguish between these alternatives, with AGCT. We compared the SGC to 100000 HGCs, used SD as the distance metric, and included stop codons in calculations. We examined six properties: hydrophobicity, polar requirement, isoelectric point, bulkiness, surface area access and fraction of accessible area lost when a protein folds. These six properties had been used previously to measure the cost for shuffling amino acid identifies (Higgs, 2009). We used AGCT to analyze each property independently, obtain an SD score for each, then covert this to a standard score. A unified score was created for each genetic code by summing all six standard scores. Rankings were determined on this basis of unified scores.

We found that, whereas the SGC performed well when compared to its HGC counterparts, the resulting unified score was no more remarkable in error minimization terms than was the summed standard score (**Figure 6**). The code ranked 105 among 100000 codes. Among the six variables, two standard scores (polar requirement and fraction of accessible area lost when a protein folds) ranked consistently lower than the unified score while the rest ranked consistently higher. The SGC ranking seems to result from two strongly conserved attributes compensating for four less-conserved attributes rather than the six properties interacting to yield a globally lowest ranking. This property combination in the SGC is neither less nor more effective at minimizing errors than is the property combination among the HGCs. Consequently, this analysis provides no evidence to suggest that the code is organized according to a negative interaction. We note, however, that negative interactions might exist. Implementing 0 for stop codon property values and standardization (*i.e.*, using standard scores and defining unified scores as sums over standard scores) may mask and temper negative interactions. Negative interactions additionally may become conspicuous in error measure analyses only after properties have been weighted according to their relative importance in protein folding. Our negative results indicate that further research considering these issues should be conducted.

Chapter 3

3.1 Testing Theories of the Genetic Code

In addition to expanding on previous research, we explored two theories that attempt to account for SGC structure. We compared the coevolution (Wong, 1975) and four column (Higgs,

2009) theories by evaluating buffering in the standard genetic code (SGC) relative to hypothetical genetic codes (HGCs) when accounting for constraints imposed by the two theories.

Both theories suggest that a primordial genetic code (PGC) preceded the SGC and the ability for the SGC to buffer effects from transition or translation error can be explained at least partially by mechanisms other than natural selection. The hypothesized PGC probably contained fewer amino acids, and additional amino acids were incorporated into the code over time (Crick, 1968; Wong, 1975; Higgs, 2009). Both theories suggest that constraints became imposed when codon positions corresponding to new entries were assigned: the coevolution theory suggests that new amino acids were constrained to associate with codons formerly associated with biosynthetic antecedents (Wong, 1975), whereas the four column theory suggest that new amino acids were constrained to associate with codons previously assigned to most-similar amino acids (Higgs, 2009). Both theories suggest that these constraints explain some or all buffering abilities in the SGC.

One way to test for buffering in the SGC is to restrict shuffling among codon block identities to positions that are hypothesized to have shared a parent codon block in the PGC. For example, in evaluating the coevolution theory, amino acid identity shuffling should be restricted to codon blocks involved in biosynthetic groups (see Freeland & Hurst, 1998b; Gilis *et al.*, 2001; Goodarzi *et al*., 2004); in evaluating the four column theory, amino acid identity shuffling should be restricted to codon blocks within columns (*i.e*., within triplets containing identical second position nucleotides). If the SGC were to rank modestly in HGC distributions generated under these constraints, then confidence in the theories would increase (*i.e*., as the observed error minimization could have been achieved under these constraints); if, however, the SGC were to rank low, then confidence in the theories would decrease.

We ran three analyses, involving 10000 HGCs and SD as the distance metric. We analyzed polar requirement for its prominence in the literature, which would allow comparisons between results obtained in the current simulations with those performed by other researchers in the past (*e.g.*, Alff-Steinberg, 1969; Haig & Hurst, 1991; Szathmary & Zinteras, 1992; Goldman, 1993; Ardell, 1998; Freeland & Hurst, 1998a, 1998b; Judson & Haydon, 1999; Higgs, 2009; Tlusy, 2010; **Table 3**). Stop codon identity positions were variable and stop codons were included in error calculations. The first analysis allowed amino acid identities to be reassigned to any codon block. The remaining two analyses corresponded to constraints imposed respectively by the coevolution and four column theories. In the coevolution theory analysis, amino acid identities could assume only positions assigned to biosynthetically related codons. In the four column theory analysis, amino acid identities could assume positions only within the column to which they are assigned in the SGC (**Figure 1**).

Although constraints imposed by both theories partially nullified error minimization in the SGC, the four column theory was more effective (**Figure 10**). Among the 10000 HGCs, the SGC ranked 6 without restricting HGCs, 315 when restricting amino acid identity reassignments to within biosynthetic families, and 1877 when restricting amino acid identity reassignments to within columns. These results can be contrasted with those obtained by Freeland & Hurst (1998b), which demonstrated that, when measuring polar requirement and differentially weighting errors caused by transitions and transversions, the SGC still achieved a low ranking relative to HGCs whose amino acid shuffling was restricted to biosynthetic pathways.

We performed another analysis, in which we weighted transitional errors more than translational errors. Whereas Freeland & Hurst (1998b) performed analyses with different weightings (*i.e.*, transitions weighted anywhere from 1 to 20 times as much as transversions), we weighted transitional errors only twice as much as transversional errors. This weighting was chosen as a conservative control, because the weighting was lower than the previously identified optimal weighting (*i.e.*, transition:transversion = 3) that favoured the SGC most remarkably (Freeland & Hurst, 1998b). Because the transition bias that was present during the origin of the PGC is unknown, a range of biases should be implemented in simulation. Choosing a weighting that is neither the most nor least favourable provided the most-objective test for the two theories. A complementary, alternative approach would involve running simulations with least and greatest realistic weightings and interpreting the results as an effect interval.

Our second analysis revealed, once again, that the four column theory was most effective at error minimization (**Figure 11**). The SGC ranked 1 in the unrestricted distribution, 4 when codon shuffling was restricted within biosynthetic families, and 722 when codon shuffling was restricted within columns. In all three analyses, the SGC ranked lower when transition bias was included. We note that, whereas the four column theory remains sufficient to explain error minimization within the SGC (μ=7.21, SD=0.66, Z=1.47, p=0.071), SGC ranking among the biosynthetically related HGC distribution is significant (μ=11.18, S.D.=2.01, Z=2.46, p=0.0069), so constraints imposed by the coevolution theory cannot explain error minimization in the SGC.

While these results increase confidence in the four column theory and fail to support coevolution theory, we cannot discount coevolution theory entirely. The results obtained for coevolution theory may have been limited by the method for generating HGCs. Describing completely the restrictions imposed by the coevolution theory would require more than simply restricting amino acid identity shuffling within biosynthetic families. We conceptually followed Di Giulio & Medugno (2000), who grouped amino acids into families based on their direct or indirect origin in the Krebs Cycle (*e.g.*, alanine, valine and leucine are members in the pyruvate family because they are either direct products in that step in the cycle or derived from amino acids that are direct products). However, this incompletely accounts for relationships within families. For instance, although the asparagine biosynthetic family contains six members (Figure 1a in Taylor & Coates, 1989; Table 1 in Di Giulio & Medugno, 2000), these six amino acids cannot be assigned to any position relative to one another. Additionally, threonine should inhabit codons adjacent to isoleucine because isoleucine is derived from threonine; this constraint is unaccounted for in AGCT. Incorporating additional restrictions such as these might raise the SGC ranking even more. Whereas results obtained with AGCT are valuable insofar as they demonstrate that co-evolution theory can explain error minimization in the SGC, additional refinement might help provide more-complete tests related to that theory.

References

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

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

Ardell, D.H. & Sella, G. (2001). On the evolution of redundancy in genetic codes. *Journal of Molecular Evolution*, 53: 269–281.

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

Balasubramanian, R., Seetharamulu, P., Raghunathan, G.A. (1980). Conformational rational for the origin of the mechanism of nucleic acid-directed protein synthesis of 'living' organisms . *Origin of Life*, 10: 15–30.

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

Crick, F.H.C. (1966). Codon-anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, 19: 548–555.

Crick, F.H.C. (1968). The origin of the genetic code. *Journal of Molecular Biology*, 38: 376–379.

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. (2005). The origin of the genetic code: Theories and their relationship, a review. *BioSystems*, 80: 174–184.

Di Giulio, M. (2008). An extension of the coevolution theory of the origin of the genetic code.. *Biology Direct*, 3: 37-58.

Di Giulio, M., Capobianco, M.R. & Medugno, M. (1994). On the optimization of the physiochemical distances between amino acids in the evolution of the genetic code. *Journal of Theoretical Biology*, 168(1): 43–51.

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.

Epstein, C.J. (1966). Role of the amino acid code and of selection for conformation in the evolution of proteins. *Nature*, 210: 25–28.

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. & Hurst, L.D. (1998b). Load minimization of the genetic code: History does not explain the pattern. *Proceedings: Biological Sciences*, 264(1410):2111–2119.

Freeland, S.J., Knight, R.D., Landweber, L.F. & Hurst, L.D. (2000). Early fixation of an optimal genetic code. *Molecular Biology and Evolution*, 17(4): 511–518.

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

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

Goldberg, A.L. & Wittes, R.E. (1966). Genetic code: Aspects of organization. *Science*, 153(3734): 420–424.

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

Goodarzi, H., Nejad, H.A. & Torabi, N. (2004). On the optimality of the genetic code, with consideration of termination codons. *BioSystems*, 77: 163–173.

Grohima, M.M., Oobatake, M. & Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry*, 82: 51–67.

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

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.

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., & Pudritz, R. E. (2007). From protoplanetary disks to prebiotic amino acids and the origin of the genetic code. *Planetary systems and the origins of life*, *3*, 1-29.

Higgs, P. G., & Pudritz, R. E. (2009). A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*,*9*(5), 483-490.

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.

Knight, R.D., Freeland, S.J. & Landweber, L.F. (2001). Rewiring the keyboard: The evolvability of the genetic code. *Nature Reviews Genetics*, 2: 49–48.

Kimura M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624-626.

Maeshiro, T. & Kimura, M. (1998). The role of robustness and changeability on the origin and evolution of genetic codes. *Proceedings of the National Academy of Sciences U.S.A.*, 95(9): 5088–5093.

Majerfeld, I. & Yarus, M. (1998). Isoleucine: RNA sites with associated coding sequences. *RNA*, 4: 471–478.

Majerfeld, I. & Yarus, M. (1994). An RNA pocket for an aliphatic hydrophobe. *Natural Structural Biology*, 1: 287–292.

Majerfeld, I., Puthenvedu, D. & Yarus, M. (2005). RNA affinity for molecular L-Histidine; Genetic code origins. *Journal of Molecular Evolution*, 61(2): 226–235.

Mannironi, C., Scerch, C., Fruscoloni, P. & Tocchini-Valentini, G.P. (2000). Molecular recognition of amino acids by RNA aptamers: The evolution into an L-tyrosine binder of a dopamine-binding RNA motif. *RNA*, 6: 520–527.

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

Milman, H.A. & Cooney D.A. (1979). Partial purification and properties of L-Asparagine synthetase from mouse pancreas. *Biochem Journal*, 181: 51–59.

Novozhilov, A.S., Wolf, Y.I. & Koonin, E.V. (2007). Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape.

Osawa, S. & Jukes, T.H. (1989). Codon reassignment (codon capture) in evolution. *Journal of Molecular Evolution*, 28: 271–278.

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

Pelc, S. R., & Welton, M. G. (1966). Stereochemical relationship between coding triplets and amino-acids. *Nature*, *209*(5026), 868-870.

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.

Szathmary, E. (1991). Codon swapping as a possible evolutionary mechanism. *Journal of Molecular Evolution,* 32: 178–182.

Szathmary, E. & Zintzaras, E. (1992). A statistical test of hypotheses on the organization and origin of the genetic code. *Journal of Molecular Evolution*, 35: 185–189.

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.

Tanford, C. (1962). Contributions of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society*, 84(22): 4240–4247.

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

Tlusty, T. (2010). A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. *Physics of Life Review,* 7(3): 362–376.

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.

Vetsigian, K., Woese, C.& Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences*, 103(28): 10696–10701.

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.

Woese CR, Kandler O, Wheelis ML. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences USA*: 87:4576-4579.

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. (1980). Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proceedings of the National Academy of Sciences USA,* 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.TF. (2005). Coevolution theory at the age of thirty. *BioEssays,* 27: 416–425.

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

Yarus, M. (1991). An RNA-amino acid complex and the genetic code. *New Biologist*, 3: 183–189.

Yarus, M. (1993). An RNA-amino acid affinity. *The RNA World*. Cold Springs Habours:

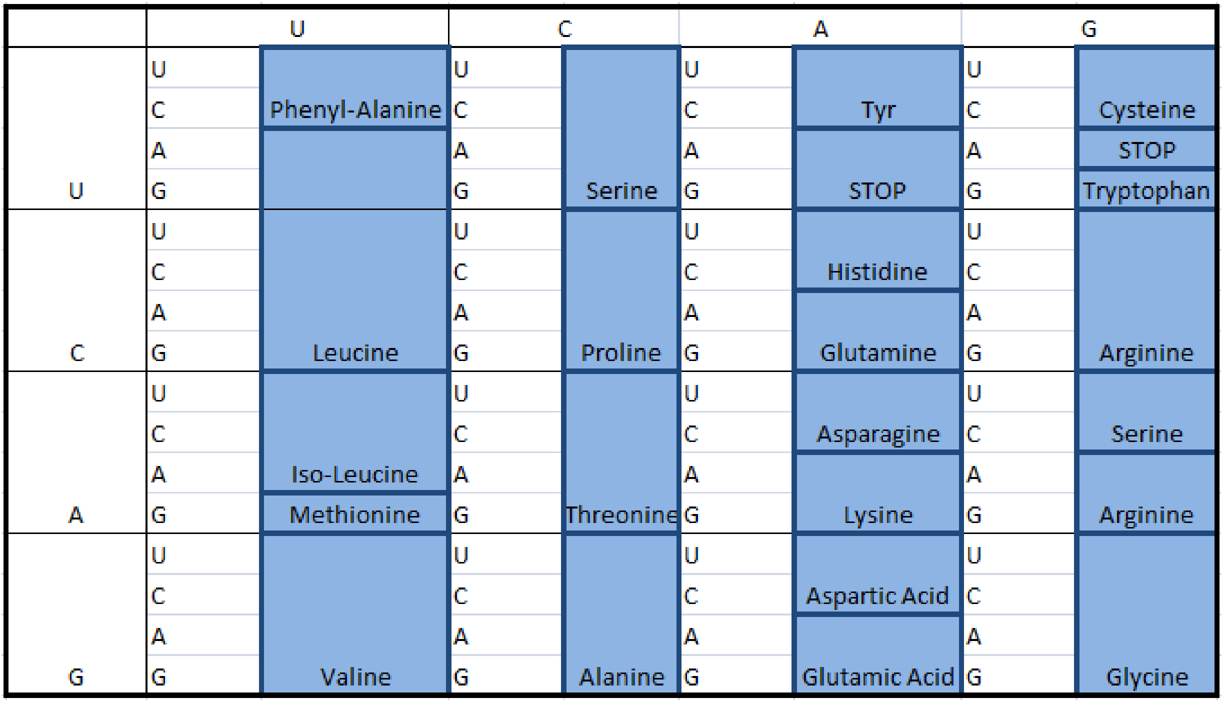
Yarus, M. (1998). Amino acids as RNA ligands: a direct-RNAtemplate theory for the genetic code's origin. *Journal of Molecular Evolution*, 47: 109–117.

Yarus, M. (2000). RNA-ligand chemistry: A testable source for the genetic code. *RNA*, 6(4): 475–484.

Yarus, M., Widmann, J.J. & Knight , R. (2009). RNA - Amino Acid bindings: A stereochemical era for the genetic code. *Journal of Molecular Evolution*, 69(5): 406–429.

Yarus, M. & Christian, F. (1989). Genetic code origins. *Nature*, 342: 349–350.

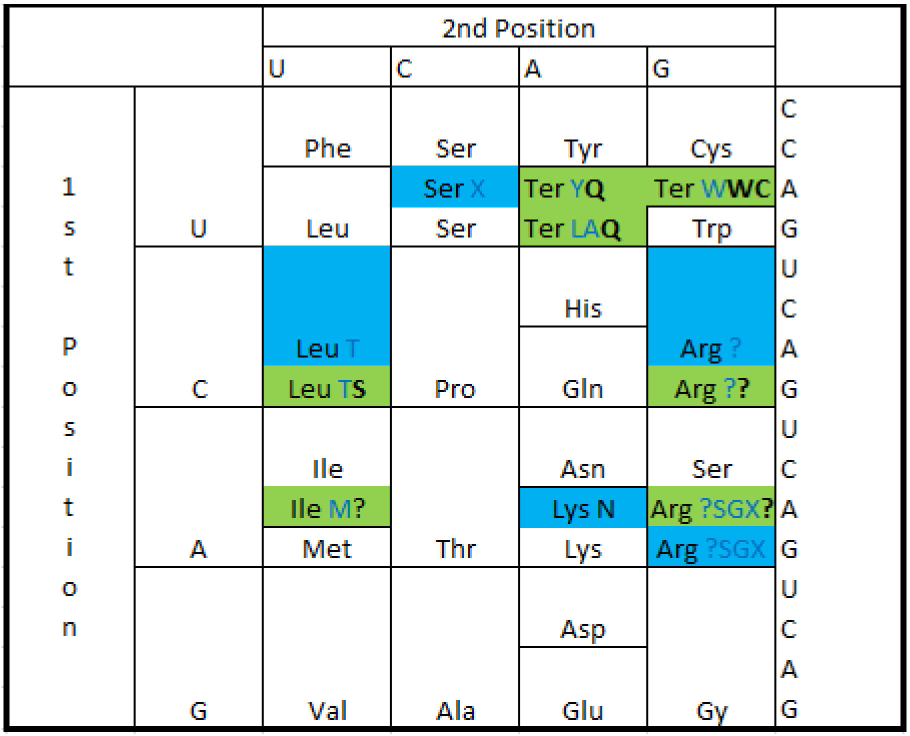
Figure 1



	U		C		A		G	
U	U	Phenyl-Alanine	U	Serine	U	Tyr	U	Cysteine
	C		C		C		C	
	A		A		A	STOP	A	STOP
	G		G		G		G	Tryptophan
C	U	Leucine	U	Proline	U	Histidine	U	Arginine
	C		C		C		C	
	A		A		A	Glutamine	A	
	G		G		G		G	
A	U	Iso-Leucine	U	Threonine	U	Asparagine	U	Serine
	C		C		C		C	
	A		A		A	Lysine	A	Arginine
	G	Methionine	G		G		G	
G	U	Valine	U	Alanine	U	Aspartic Acid	U	Glycine
	C		C		C		C	
	A		A		A	Glutamic Acid	A	
	G		G		G		G	

The standard genetic code in canonical presentation.

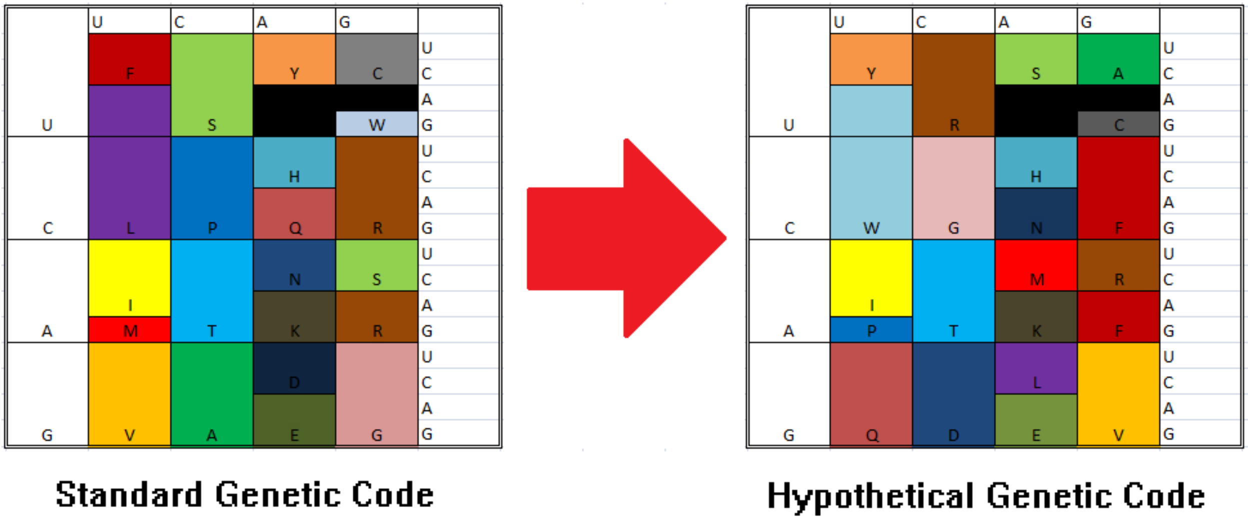
Figure 2



		2nd Position				
		U	C	A	G	
						C
		Phe	Ser	Tyr	Cys	C
1			Ser X	Ter YQ	Ter WWC	A
s	U	Leu	Ser	Ter LAQ	Trp	G
t						U
				His		C
P		Leu T			Arg ?	A
o	C	Leu TS	Pro	Gln	Arg ??	G
s						U
i		Ile		Asn	Ser	C
t		Ile M?		Lys N	Arg ?SGX?	A
i	A	Met	Thr	Lys	Arg ?SGX	G
o						U
n				Asp		C
						A
	G	Val	Ala	Glu	Gy	G

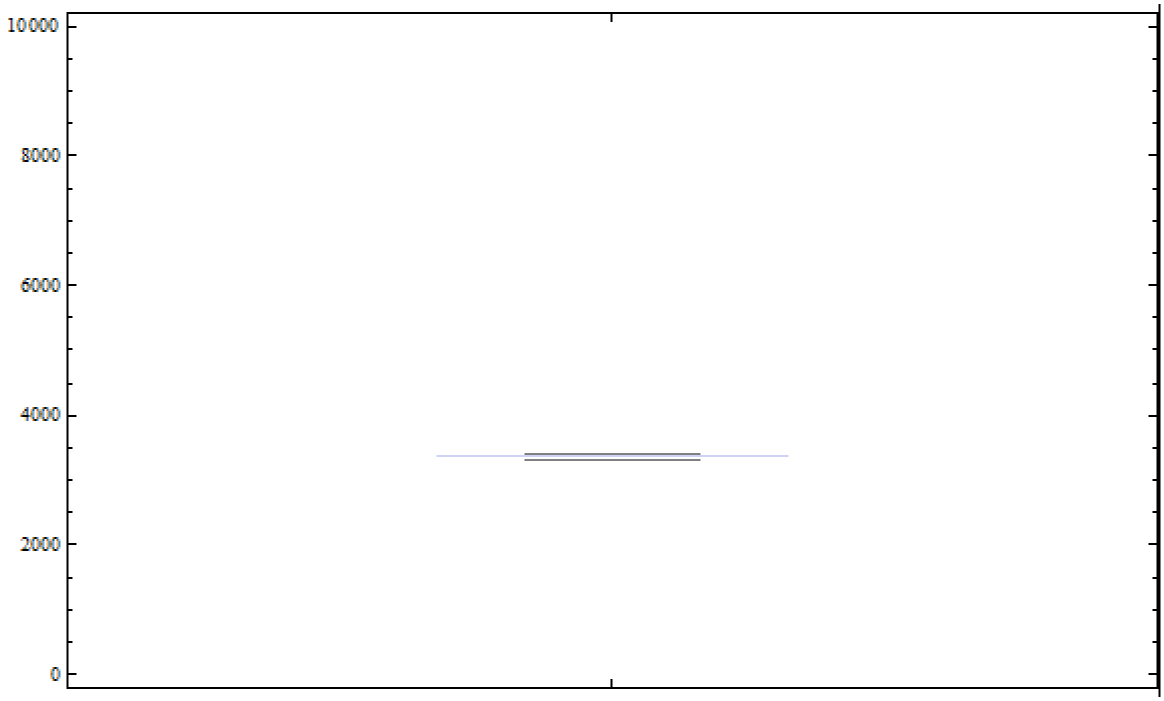
Changes to standard genetic code (adapted from **Figure 3** in Knight *et al.*, 2001). Blue shading represents codon blocks that change in mitochondrial lineages; green shading represents codon blocks that change in mitochondrial and nuclear lineages. Individual letters within codon blocks are standard one-letter amino acid abbreviations (*i.e.*, T=threeonine, M=methionine, S=serine, Y= tyrosine, L= leucine, Q=glutamine, A=alanine, W=tyrptophan, C=cysteine, G= glycine, N=asparagine, X=unknown, ?=unassigned). Blue letters indicate changes that have occurred in mitochondrial lineages, whereas black letters indicate changes that have occurred in nuclear lineages. Question marks represent stop codons.

Figure 3



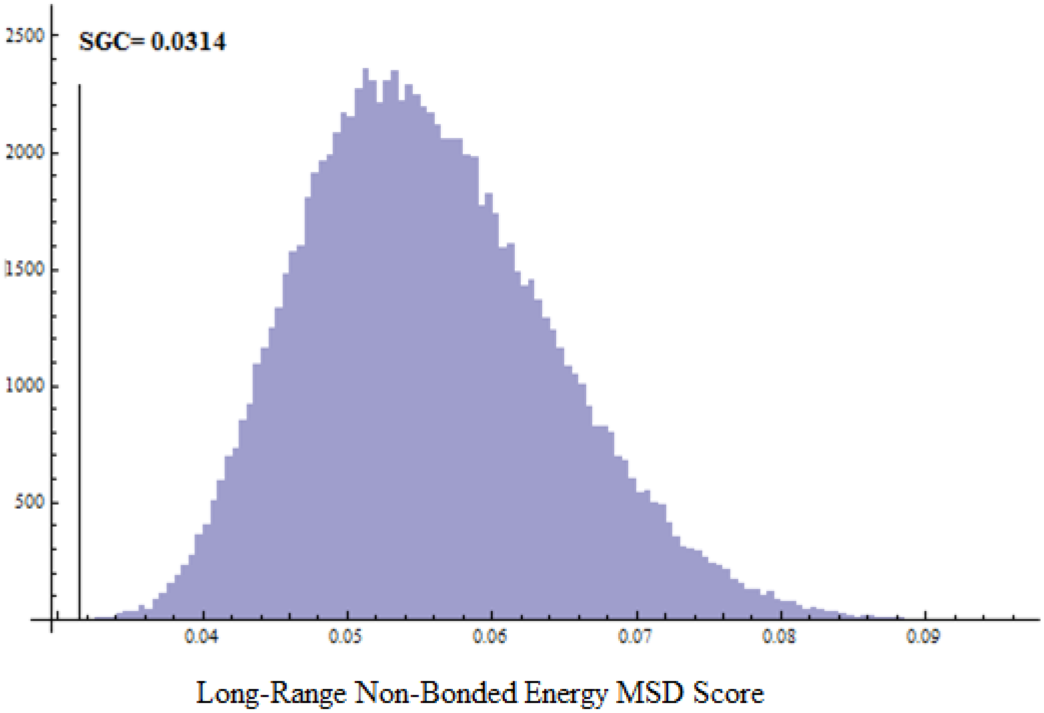
The standard genetic code may be transformed into a hypothetical genetic code by shuffling amino acid identities (indicated by colour and letter).

Figure 4



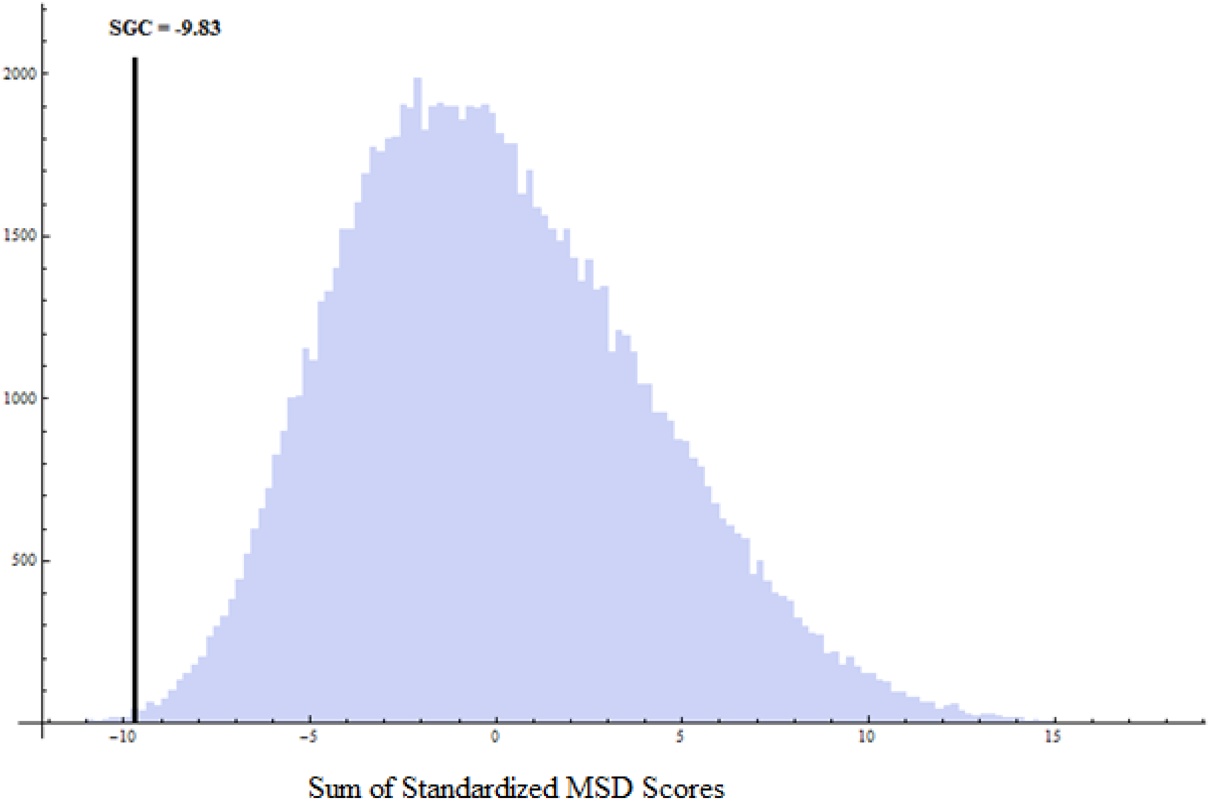
A box plot representing the ranking for the standard genetic code amid 10000 hypothetical genetic codes (mean ≈ 3361, horizontal line; standard deviation ≈ 31, shaded rectangle) using mean squared distances to calculate effects from potential transcription and translation errors on hydrophobicity.

Figure 5



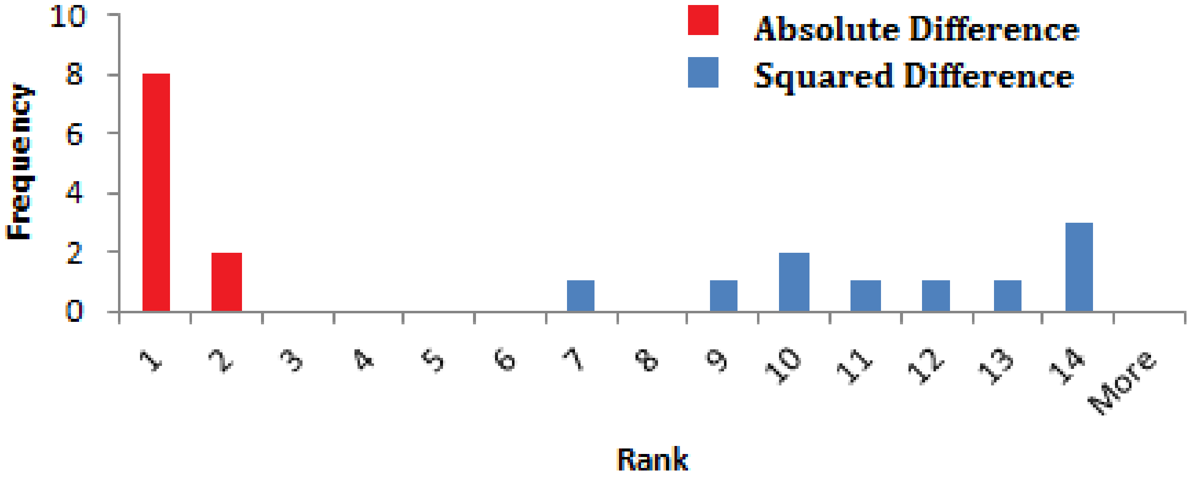
Ranking for the standard genetic code relative to 100000 hypothetical genetic codes for mean square error in long-range non-bonded energy. The standard genetic code ranked 2, meaning that only one hypothetical genetic code was more buffered.

Figure 6



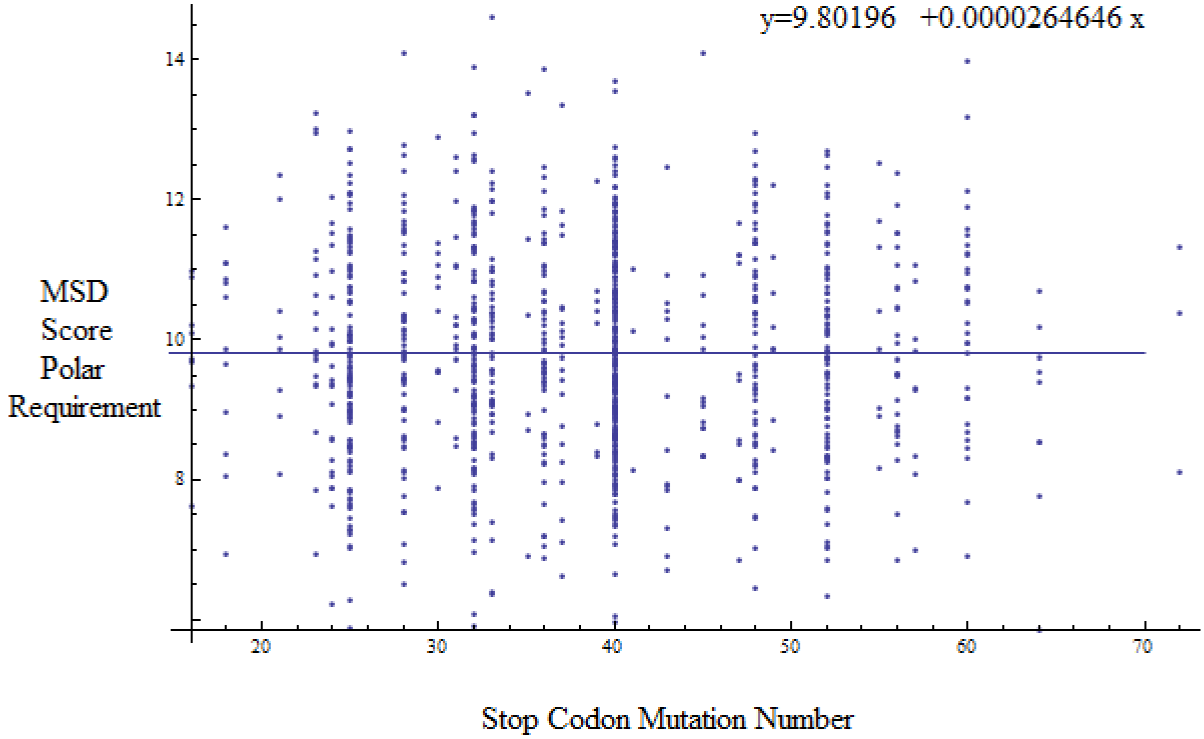
Ranking for the standard genetic code relative to 100000 hypothetical codes for mean square error over six properties: hydrophobicity, polar requirement, isoelectric point, bulkiness, surface area access and fraction of accessible area lost when a protein folds and polar requirement. The standard genetic code ranked 105.

Figure 7



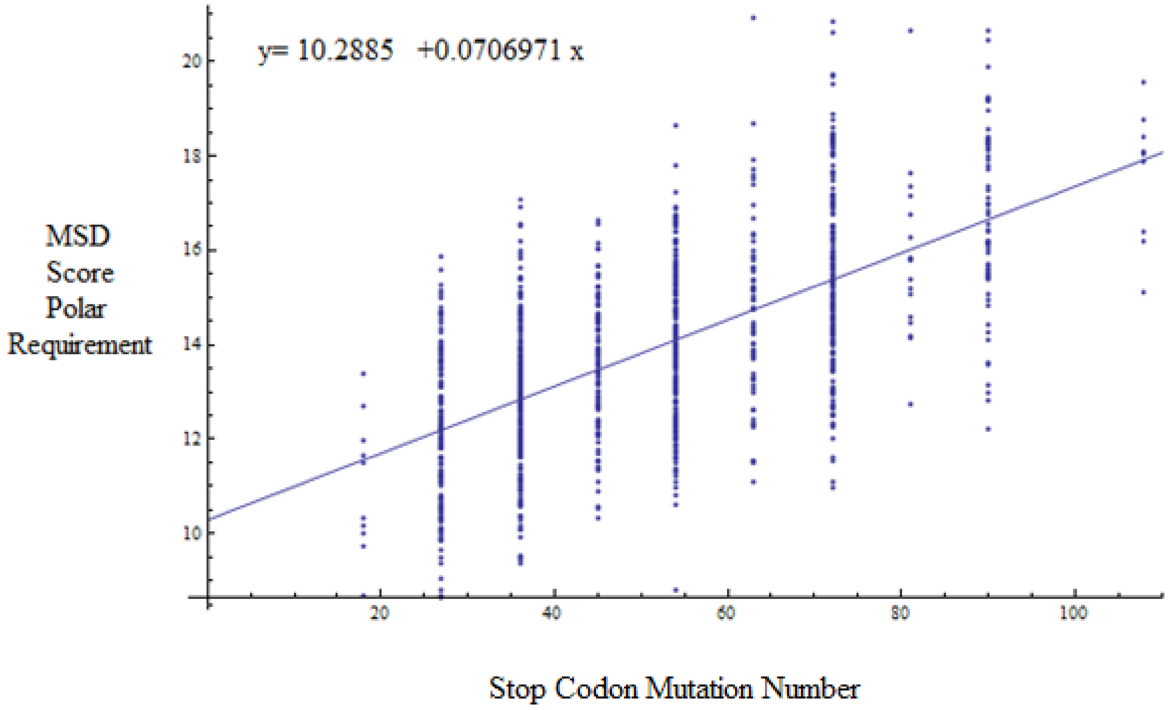
Ranking for the standard genetic code relative to 10000 hypothetical genetic codes using unified scores (*i.e.*, standard scores summed) over six properties: hydrophobicity, polar requirement, isoelectric point, bulkiness, surface area access, fraction of accessible area lost when a protein folds and polar requirement. Results for two different distance metrics are shown.

Figure 8



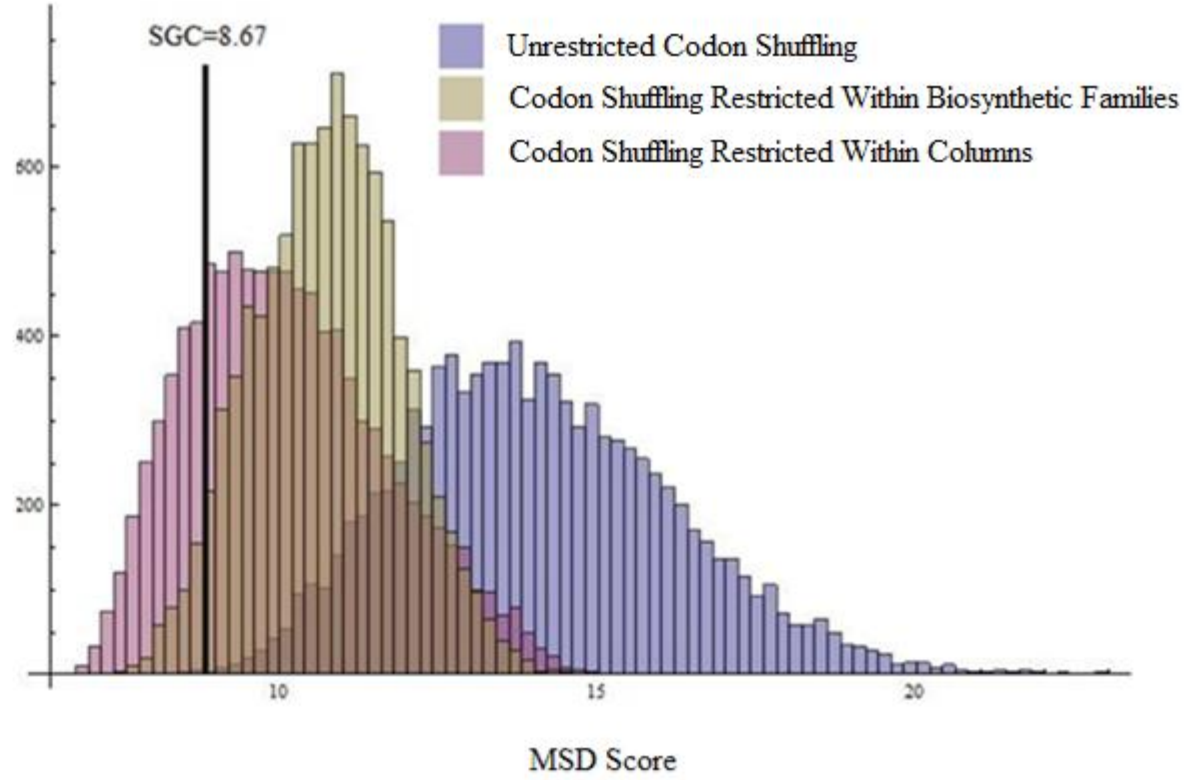
Plot showing mean square distance polar requirement scores (ordinate) and stop codon mutation numbers (c) for 1000 hypothetical genetic codes, with best-fit regression line equation. Stop codons were excluded from error measure calculations.

Figure 9



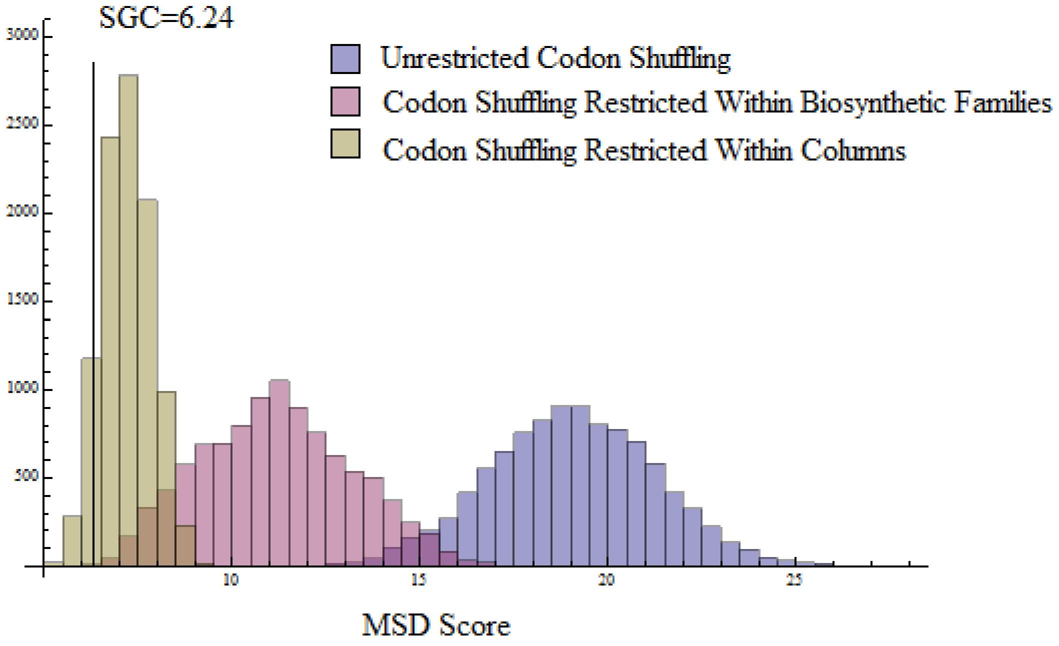
Plot showing mean square distance polar requirement scores (ordinate) and stop codon mutation numbers (abscissa) for 1000 hypothetical genetic codes, with best-fit regression line equation. Stop codons were included in error measure calculations.

Figure 10



Ranking for the standard genetic code relative to three 100000 hypothetical genetic code populations for mean square distance in polar requirement. Unrestricted codon shuffling allowed amino acid identities to be shuffled among any codon position, whereas the other conditions involved shuffling restricted on the basis of coevolution theory and the four column theory.

Figure 11



Ranking for the standard genetic code relative to three 100000 hypothetical genetic code populations for mean square distance in polar requirement. Unrestricted codon shuffling allowed amino acid identities to be shuffled among any codon position, whereas the other conditions involved shuffling restricted on the basis of coevolution theory and the four column theory. Errors attributed to transitions were weighted twice as much as were errors attributed to transversions.

Table 1

Serine Family	Phosphoenolpyruvate Family	Pyruvate Family	Aspartate Family	Gultamate Family
Serine Tryptophan Cysteine Glycine	Phenyl-Alanine Tyrosine	Alanine Valine Leucine	Aspartate Asparagine Threeonine Isoleucine Methionine Lysine	Glutamic Acide Glutamine Arginine Proline Histidine

Biosynthetic families published by Wong (1975).

Table 2

Phosphoenolpyruvate Family	Pyruvate Family	Aspartate Family	Gultamate Family	Misc. Family
Phenyl-Alanine Tyrosine	Alanine Valine Leucine	Aspartate Asparagine Threeonine Iso-Leucine Methionine Lysine	Glutamic Acide Glutamine Arginine Proline	Histidine

Biosynthetic families published by Di Giulio & Medugno (2000).

Table 3

Paper	Property																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Alff-Steinberger, 1969	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0
Haig & Hurst, 1991	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
Szathmary & Zinteras, 1992	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Goldman, 1993	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Freeland & Hurst, 1998a	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Freeland & Hurst, 1998b	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ardell, 1998	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Judson & Haydon, 1999	1	1	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0
Gilis *et al.* , 2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Higgs, 2009	0	1	0	0	1	0	1	0	1	0	0	1	1	1	1	0	0
Tlusy, 2010	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

A	Molecular Weight
B	Polar Requirement
C	Number of Dissociating Groups
D	PKi
E	Isoelectric Point
F	Alpha-Helix Forming Ability
G	Molecular Volume
H	Hydropathy
I	Bulkiness
J	Composition
K	Refractivity
L	Hydrophobicity Scale
M	Surface Area Accessible to Water when Unfolded
N	Surface Area Lost When Protein Folds
O	Polarity
P	Torsion Potential
Q	Long-Range Non-Localized Hydrophobic Interactions

Amino acid properties examined in previously published papers. A '1' indicates the property was analyzed, whereas a '0' indicates that a property was unanalyzed.

Table 4

Compressibility	K^0
thermodynamic transfer hydrophobicity	H_{t}
surrounding hydrophobicity	$H\mathrm{p}$
Polarity	P
isoelectric point	PH_{i}
equilibrium constant with reference to the ionization property of cooh group	$\mathrm{P}K'$
molecular weight	M_{w}
Bulkiness	B_{l}
chromatographic index	R_{f}
refractive index	M
normalized concensus hydrophobicity	H_{nc}
short and medium range non-bonded energy	E_{sm}
long-range non bonded energy (london forces + electrostatic)	E_l
total non-bonded energy	E_t
alpha helix tendency	P_{α}
b structure tendency	P_{β}
turn tendency	P_{t}
coil tendency	P_{c}
helical contact area	C_{a}
rms fluctuational displacement	F
Buriedness	B_{r}
solvent accessible reduction ratio	R_{a}
average number of surrounding residues	N_{s}
power to be at the n terminal	α_{n}
power to be at the c terminal	α_{c}
power to be at the middle of an alpha helix	α_{m}
partial-specific volum	V^0
average medium contacts	N_{m}
long range contacts (inter molecular stabalization)	N_{l}
combined surrounding hydrophobicity	H_{gm}
solvent accessible surface area for denatured	ASA_{D}
solvent accessible surface area for native	ASA_{N}
solvent accessible surface area for unfolding	ΔASA
gibbs free energy change of hydration for unfolding	ΔG_{h}
gibbs free energy change for denatured	G_{hD}
gibs free energy change for native	G_{hN}
unfolding enthalpy change of hydration	ΔH_{h}

untfolding enthalpy change of hydration	$-T\Delta S_{\mathrm{h}}$
unfolding hydration heat capacity change	ΔC_{ph}
unfolding gibbs free energy	ΔG_{c}
unfolding enthalpy	ΔH_{c}
unfolding enthalpy change of hydration	$-T\Delta S_{\mathrm{c}}$
gibbs free energy change	ΔG
unfolding enthalpy	ΔH
unfolding enthalpy changes of the chain	$-T\Delta S$

The 45 properties published by Grohima *et al.* 1999 and their respective symbols, analysed herein.

Table 5

The 49 properties examined through AGCT and the associated ranking for the standard genetic code.

Property	Rank	Property	Rank	Property	Rank	Property	Rank	Property	Rank	Bin	Color
E_t	1	E_l	125	hydrophobicity	254	V^0	2299	E_{sm}	8247	Strongly Conserved (<100)	🟩
H_{nc}	2	F	172	ΔG	332	melting Point	2354	N_m	8401	Conserved (<250)	🟩
$-T\Delta S_c$	3			ΔH	566	P	2840	G_{hN}	8795	Non-Conserved Low (250<x<2000)	🟨
polar requirement	16			P_β	580	α_n	2884			Non-Conserved Medium (2000<=x<=8000)	🟥
N_l	19			µ	682	P_t	3202			Non-Conserved High (>8000)	🟥
ΔH_c	23			$-T\Delta S_h$	710	P_α	3254				
R_f	28			H_t	972	C_a	3684				
ΔASA	29			$-T\Delta S$	996	B_l	3801				
ΔC_{ph}	41			ASA_N	1237	ASA_D	4093				
N_s	70			ΔG_c	1254	ΔG_h	4669				
R_a	79			K^0	1393	absolute Entropy	4704				
B_r	91			M_w	1600	α_m	4739				
H_p	99			P_c	1997	H_{gm}	5301				
						ΔH_h	5543				
						α_c	6188				
						PH_i	7011				
						PK'	7464				
						G_{hD}	7521				