Emergence of the Canonical Genetic Code

Fall 2009 Emergent States of Matter Term Essay

Samuel O. Skinner

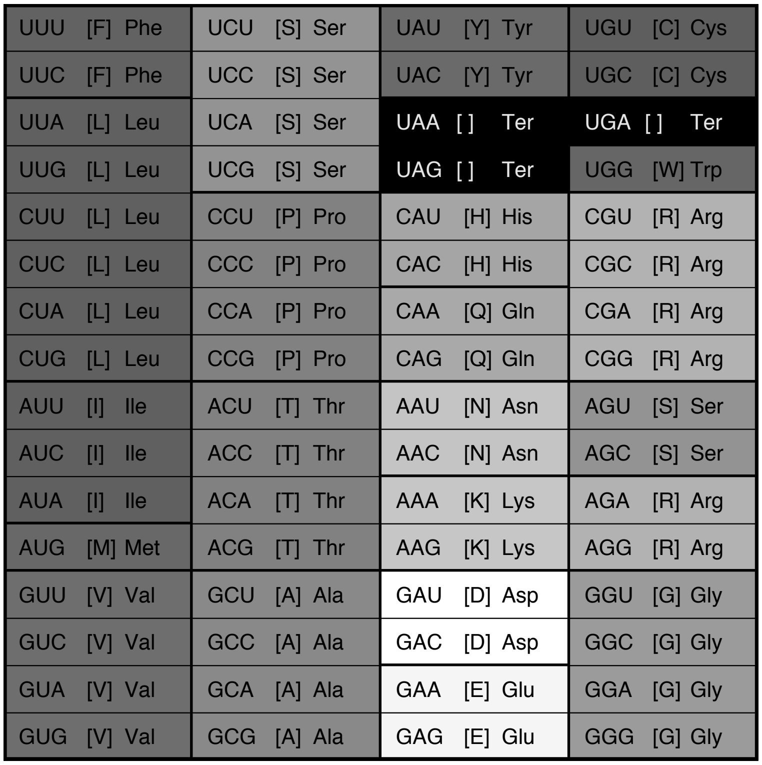
Abstract

I will review literature that expands the theory that the canonical genetic code was not universal in early life. Instead, the rudimentary genetic code and decoding mechanism were highly ambiguous, in that mistranslations were made and tolerated. From this early communal state, genetic codes diverged in pools of innovation. Through the interaction and competition of these pools, the optimization and universality of the canonical genetic code emerged.

Introduction

All organisms survive by utilizing the information encoded in sequences of nucleotides. The translation of this information into functional proteins requires a means of interpretation, which consists of the genetic code and the de-coding mechanism. The genetic code is the mapping from the 64 triplets of nucleotides (codons) to the 20 naturally occurring amino acids as well as sequences that indicate a 'start' or 'stop' of translation. When a desired sequence of nucleotides is translated by the de-coding mechanism, the mapped sequence of amino acids will produce the desired protein that is able to carry out a particular function of the organism.

After the genetic code was deciphered through the study of *Escherichia coli* (Nirenberg, Clark et al. 1963), it was recognized that the code is universal to all life (Crick 1968). Recently, variants in the genetic code have been documented, but they are all believed to be derived from the canonical genetic code (Knight, Freeland et al. 2001). Inspections of the genetic code quickly showed that it has nonrandom structure (Figure 1). A property of the code that is easily recognized is the block structure, where neighboring codon sequences (sequences differing by one nucleotide) are assigned to the same or similar amino acids. Almost all proteins we see today are highly specialized,



UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [ ] Ter	UGA [ ] Ter
UUG [L] Leu	UCG [S] Ser	UAG [ ] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

Figure 1 | The Universal and Optimized Genetic Code (Each box contains nucleotide sequence and amino acid (in two abbreviations).) The table displays the mapping from 64 nucleotide triplets to 20 amino acids and start (AUG, also coding for Met) and stop (UAA, UAG, UGA - Ter). This code is (with very few exceptions) universal to all life on earth. Described below are examples of the optimality of the genetic code.

The block structure is easily noticeable, where codons differing by one amino acid tend to be assigned to the same amino acid. This is thought to minimize the effect of translation errors and mutations in the nucleotide sequence on the translated amino acid sequence. Because, if a mistranslation event occurs, there is a chance a codon will be mistaken for a synonymous codon. Similarly, codons differing by one amino acid tend to be assigned to similar amino acids. This similarity has been best represented by an amino acid's "polar requirement" (Woese, Dugre et al. 1966), corresponding to the shading in the above table.

Image from (Koonin and Novozhilov 2009).

to a point where substitution of an amino acid would most likely have deleterious effects to the functionality of the protein. Amino acid substitutions due to translation errors or genome mutations would most likely result with the amino acid of a neighboring codon. So, the redundancy and neighbor similarity in the block structure has been accepted to minimize deleterious effects of amino acid substitutions (Woese 1965; Woese, Dugre et al. 1966). To investigate the optimality of the canonical genetic code with respect to the cost of translation error, comparisons were made against randomly generated codes. Though there is not a universal measure that assesses how translation errors are deleterious to the organism, it is accepted that the canonical genetic code is optimized to a surprising extent (Haig and Hurst 1991; Freeland and Hurst 1998; Novozhilov, Wolf et al. 2007; Butler, Goldenfeld et al. 2009).

The discovery of the universality and optimality of the genetic code only promotes questions on its origin. The origin of the code has been postulated to be: (1) a ‘frozen accident’, where after the 20 amino acids had been incorporated into the code any changes would be lethal (Crick 1968), (2) a result of stereo-chemical interactions of the amino acids and codons, (3) a result of selection for translation-error minimization (Woese 1965; Koonin and Novozhilov 2009), and (4) the result of co-evolution of the code with the synthesis of novel amino acids (Higgs 2009). The code’s recently calculated optimality seems to suggest a period of evolution (Butler, Goldenfeld et al. 2009). The above theories are not mutually exclusive, or exclusive with respect to a period of code evolution. Additionally, recent examples of code variants (believed to be derived from the canonical code) suggest that the code is not frozen and still evolving (Knight, Freeland et al. 2001). There is still the need for an experimentally based or rigorous explanation for the observed universality and optimality.

The universality of the code suggests that the period of significant evolution would have been during the time of the Last Universal Common Ansestor (LUCA). A suitable theory of the code’s evolution also requires a suitable understanding of the nature of this time. Phylogenic studies have uncovered the prevalence of horizontal gene transfer (HGT) in early life. In this time, the evolutionary dynamic of organisms was significantly different that Darwinian (or, ‘vertical’) evolution. Instead of speciation, a communal state evolved as biological innovations were shared (Woese 1998). With the knowledge of such a different evolutionary dynamic during the period of early life and before LUCA, new theories can be shaped about how the genetic code came to be.

Hypotheses for the code’s evolutionary dynamic

Recently, it has been postulated that HGT provided evolutionary pressure that produced the universality and optimization of the genetic code.

It has been suggested that the precision of the genetic code and decoding mechanism coevolved from a rudimentary communal state. The communal state would encourage the use of shared protocols, or genetic codes. Through the interaction of communities using different protocols, the codes would converge. The central hypothesis investigated is that the universality of the genetic code may have been a necessary condition for life to evolve into the complex state capable of undergoing a Darwinian transition, and that horizontal gene transfer is the interaction mechanism by which the universal genetic code emerged (Syvanen 2002; Vetsigian, Woese et al. 2006). Vetsigian et al 2006, pursued this hypothesis using computational simulations to predict the code's evolution, through a dynamic of innovation sharing. This paper predicts that this dynamic inevitably leads to the observed universality and optimization. The predictions of these simulations will be the focus of this review.

Since the 1960's, it has been fairly postulated that the progenote is a most rudimentary organism, containing simplified versions of the translation machinery that we know to currently be an intricate complex of macromolecules (Woese 1965; Woese 1998; Vetsigian, Woese et al. 2006). A simplified version of the translation machinery is argued to necessarily be less accurate in the assignment of nucleotide sequences to amino acid sequences. However, this ambiguity would be tolerated and embraced in early life. Sets of codons would be translated to sets of amino acid sequences, giving rise to the concept of "statistical proteins". The communal state of these organisms would contain very high mutation and genetic exchange rates, where "essential functions" such as DNA replication, translation, and protection from genetic exchange were still being developed. So, innovations made by chance could be globally distributed, which shapes the evolutionary dynamic of this period of early life (Woese 1965; Woese 1998).

The sharing of innovation requires a common protocol, or a shared genetic code. Cells that use the same genetic code will be able to share genetic material freely. A large community of cells would be able to produce niches that have specialized in diverse ways. Thus, a large and diverse community using the same genetic code would be able to produce and share more innovations. When communities of cells with different genetic codes come into contact, genetic material can be transferred through HGT, but a process of conversion must take place for the incorporation of the innovation. The host's translation machinery must undergo a "detuning" process to attempt to accept the transferred gene. It is beneficial for the foreign gene to be shifted into the host's code (references for detuning processes in Vetsigian et al. 2006). The result would be the incorporation of the innovation into the host and the slight modification of the host's code to be more able to interpret the innovations of the new community. Under these assumptions, the codes are attractive: the more similar the codes, the easier it is for the innovations to be incorporated and for codes to converge. Because larger communities have access to more innovations than smaller ones, the genetic code of the larger community will generally be able to out-compete that of the smaller ones. So, interactions of genetic codes through HGT are predicted to inevitably give rise to a universal code (Vetsigian, Woese et al. 2006).

In addition to the interaction between communities using different genetic codes, the transference of translation machinery will participate in the evolution of the genetic code. If the components of the translation machinery are considered to be subject to the discovery of more optimal configurations and modifications, the optimality of the code is an evolvable characteristic as well. Innovations of the translation machinery can be spread throughout communities, and can be assumed to be accepted because of the universal benefit of optimality for all genetic codes (Vetsigian, Woese et al. 2006).

The main hypothesis of these papers is that the universality and optimality of the genetic code are not strictly due to stereochemical codon-amino acid assignment or a "frozen accident". Instead, these characteristics are primarily the result of the genetic code's evolutionary dynamic shaped by the interactions of different codes through HGT. Through these interactions, the universal genetic code emerged capable of allowing the accelerated exchange of innovation ultimately resulting in the evolutionary dynamic of vertical decent.

Methods

The above hypotheses were tested using computational simulations of coevolving species each consisting of a genome and a genetic code. All of the species communally evolve, sharing innovations (in this case, beneficial outcomes of mutation and code evolution) through HGT (Vetsigian, Woese et al. 2006). The algorithm describing the co-evolution of the genome and genetic code was based on previous simulations (Sella and Ardell 2002). The principle goal is to present a mechanism that results in universality and optimality of the genetic code. The hypothesized dynamical nature of early life described above was represented in the following way.

The genomes of individual species are subject to genetic mutations and selection pressure. Each species starts with randomly assigned genetic codes. To simulate the dynamic of HGT, a fraction of each genome is replaced by foreign genetic material from random donors. The acceptor then attempts to make an incremental change to the code to make beneficial use of its newly composed genome.

The optimality of the codes was quantified by scoring the assigned amino acid similarity between related codon sequences. Simulations were done with and without horizontal gene transfer for comparison.

Results

The primary results of the simulation are that they demonstrate HGT as a mechanism that will lead inevitably to the convergence of interacting genetic codes. The probability of universality depends on the degree of HGT considered in the simulations, but it is one after a threshold HGT value. Interestingly, the simulations also demonstrate that the genetic codes initially diverged as they gained complexity and specificity. The codes (or, code in the case it is universal) are optimized to a much greater extent in the presence of horizontal gene transfer. As a comparison, the simulations performed with no horizontal gene transfer similarly diverged and were optimized to an extent, but remained disparate and consequently resulted in less optimal codes.

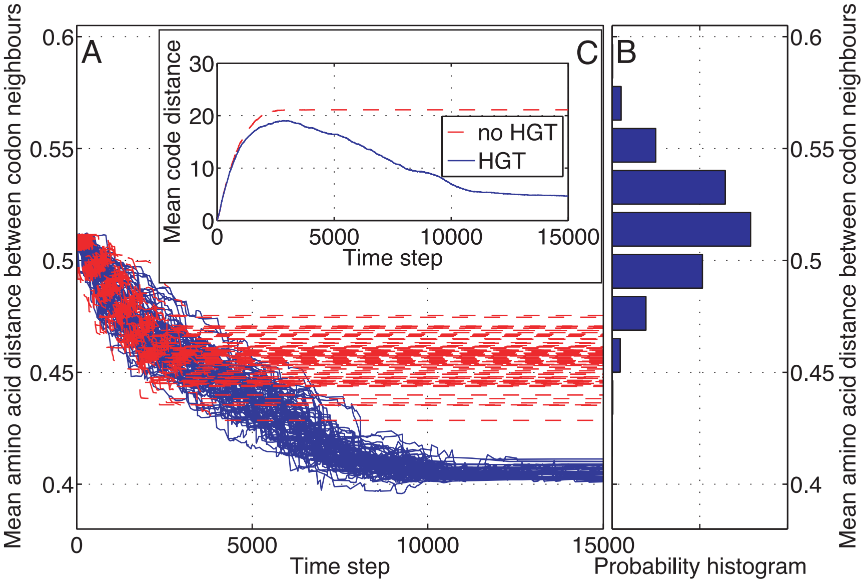


Figure 2 | Collective evolution of the genetic code due to HGT Figure from Vestigian et al. 2006. The co-evolution of genetic codes with and without HGT are shown above. Over time, genetic codes converge and optimize in the presence of HGT (blue). Without HGT, simulations of genetic codes show slight optimization, but do not converge (red). Inset shows how genetic codes initially diverge. In the presence of HGT codes will converge (blue), and codes without HGT will remain different (red). The histogram on the right is the optimality distribution of randomly generated codes. These trajectories display the emergence code universality through the interaction of HGT.

Discussion

Vetsigian et al. 2006 expands upon previous computational studies of the evolution of the genetic code, incorporating recent evidence for collective evolution before the onset of Darwian evolution. These simulations provide a basis for understanding the observed optimality and universality of the genetic code by demonstrating that they could have emerged generically from the dynamical interactions of the communal state. This paper also provides an appealing picture of the biological state before the onset of vertical evolution. After the genetic code has evolved to a universal state, horizontal gene transfer would allow exponential increases in complexity. This would lead quickly to the Darwinian transition, where vertical descent is more beneficial (Vetsigian, Woese et al. 2006).

The results of the simulations provide evidence for the predictions set by the authors. To describe the system in more detail, the authors do list mechanisms that can be incorporated in future simulations. To accurately simulate the hypothesized rudimentary progenote, one would attempt to include the translational components as evolvable mechanisms. Also, to accurately depict the theoretical environment above, the competition between codes would have to be included in the dynamics, where codes can be completely integrated into another. However, this most likely would reinforce the universality conclusion of the simulations.

These simulations did show that the genetic code was optimized to a great extent in the presence of horizontal gene transfer. Most importantly, these simulations provided a demonstration of how the code's (previously very elusive) properties can generically emerge from simple interactions. This result is consistent with the theory that the origin of the canonical genetic code is a mixture of the 'frozen accident' hypothesis and translation-error minimization. Predictions of the degree of optimization and the structure of the codon table will have to come from more specific simulations. Simulations with these aims have been pursued, incorporating the code's co-evolution with speculated non-biological amino acid formation and biosynthetic pathways (Higgs 2009). These simulations have predicted that the genetic code evolved from a base set of amino acids. Under these assumptions, the block structure is an inherent property of the codon table. When a new amino acid is incorporated, a block of a similar amino acid is sub-divided. With further detail incorporating into simulations, it seems possible to find elements of interactions from all existing theories.

The integration of co-evolution of the code and amino acid synthesis with HGT would be an interesting, and complicated, line of research. It would be interesting to see how this temporal progression of increase in amino acids would perform in a simulation with the communal dynamic in the articles discussed above. The development of the universal code seems like it would depend on how the new amino acid is placed into the codon table: either deterministically, or stochastically. In the deterministic case, where it is placed in the block of the most similar amino acid, it seems most likely that code universality would occur much more quickly. If the placement occurs stochastically, each time an amino acid is introduced it would seem that the codes would undergo a period of diversification, similar to that seen in the simulations in the surveyed paper.

Summary and Conclusion

A detailed explanation for the origin (and perhaps the current trajectory) of the genetic code is still not complete. Surprisingly, most of the theories of the origin of the genetic code were made shortly after it discovery, about 40 years ago. Mechanisms of: a 'frozen accident', stereo-chemical interactions, co-evolution of amino acids, and translation-error minimization all could play a role in how the canonical genetic code came to be universal and optimized. It seems that many questions remained because, besides the code itself, there is little evidence we could gather about early life.

However, the evidence for HGT in the deep past produced a new environment for how the canonical genetic code's optimality and universality could be explained. In Vetsigian et al. 2006, HGT was integrated into previous genetic code evolution algorithms to predict how interactions of organisms could lead inevitably to an optimized, unified protocol for sharing innovation, the genetic code. The simulations predict that interactions not involving HGT would become diverse and slightly optimized, but not to the extent of those including HGT. These findings strongly support the theory of a communal state made of rudimentary organisms, where the fundamental process such as translation and replication were still evolving. From this communal state evolved a universal code that could produce the complexity of life that we see today. However, a full understanding of the structure, optimality, and evolution of the genetic code requires much more experimental evidence.

It is becoming quite clear that understanding biological complexity requires a shift from reductionist approaches to an approach studying systems and collective phenomenon. This is demonstrated in Vetsigian et al 2006, where simulations were used to predict emergent characteristics from the interactions known to have existed in the early communal state.

References

Butler, T., N. Goldenfeld, et al. (2009). "Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement." Phys Rev E Stat Nonlin Soft Matter Phys **79**(6 Pt 1): 060901.

Crick, F. H. (1968). "The origin of the genetic code." J Mol Biol **38**(3): 367-79.

Freeland, S. J. and L. D. Hurst (1998). "The genetic code is one in a million." J Mol Evol **47**(3): 238-48.

Haig, D. and L. D. Hurst (1991). "A quantitative measure of error minimization in the genetic code." J Mol Evol **33**(5): 412-7.

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." Biol Direct **4**: 16.

Knight, R. D., S. J. Freeland, et al. (2001). "Rewiring the keyboard: evolvability of the genetic code." Nat Rev Genet **2**(1): 49-58.

Koonin, E. V. and A. S. Novozhilov (2009). "Origin and evolution of the genetic code: the universal enigma." IUBMB Life **61**(2): 99-111.

Nirenberg, M. W., B. F. C. Clark, et al. (1963). "On Coding of Genetic Information." Cold Spring Harbor Symposia on Quantitative Biology **28**: 549-&.

Novozhilov, A. S., Y. I. Wolf, et al. (2007). "Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape." Biol Direct **2**: 24.

Sella, G. and D. H. Ardell (2002). "The impact of message mutation on the fitness of a genetic code." J Mol Evol **54**(5): 638-51.

Syvanen, M. (2002). "Recent emergence of the modern genetic code: a proposal." Trends Genet **18**(5): 245-8.

Vetsigian, K., C. Woese, et al. (2006). "Collective evolution and the genetic code." Proc Natl Acad Sci U S A **103**(28): 10696-701.

Woese, C. (1998). "The universal ancestor." Proc Natl Acad Sci U S A **95**(12): 6854-9.

Woese, C. R. (1965). "On the evolution of the genetic code." Proc Natl Acad Sci U S A **54**(6): 1546-52.

Woese, C. R., D. H. Dugre, et al. (1966). "On the fundamental nature and evolution of the genetic code." Cold Spring Harb Symp Quant Biol **31**: 723-36.