What can information-asymmetric games tell us about the context of Crick's "Frozen Accident"?

Justin Jee[1,2,*], Andrew Sundstrom[1], Steven E. Massey[3], Bud Mishra[1,2,*]

[1.] Courant Institute of Mathematical Sciences, New York University, New York, NY 10003
[2.] Sackler Institute of Biomedical Sciences, New York University, New York, NY 10016
[3.] Biology Department, University of Puerto Rico - Rio Piedras, San Juan, Puerto Rico, USA 00931

[*] To whom correspondence should be addressed
mail: NYU Bioinformatics Lab, 10th Floor, 715 Broadway Ave, New York, NY 10003
phone: 212.998.3464
email: justin.jee@med.nyu.edu, mishra@nyu.edu

Abstract

This paper describes a novel application of information-asymmetric (signaling) games to molecular biology in which utility is determined by the message complexity (rate) in addition to the error in information transfer (distortion). We show using a computational model how it is possible for the agents in one such game to evolve a signaling convention (separating equilibrium) that is suboptimal in terms of information transfer, but is nonetheless stable. In the context of an RNA world merging with a nascent amino acid one, such a game's equilibrium is alluded to by the genetic code, which is nearly optimal in terms of information transfer, but is also universal and nearly immutable. Such a framework suggests that cellularity may have emerged to encourage coordination between RNA species and sheds light on other aspects of RNA world biochemistry yet to be fully understood.

The genetic code, the mapping of nucleic acid codons to amino acids via a set of tRNA and aminoacylation machinery, is near-universal and near-immutable. In addition, the code is also near-optimal in terms of error minimization, i.e. tRNAs recognizing similar codons may be mistaken for each other during translation, yet these mistakes often have no negative impact on translation because similar codons map to identical amino acids or ones with similar physiochemical properties [1][2]. Biochemists have long wondered: If immutability and universality were early properties (i.e. the genetic code was a "frozen accident" [3]), then how could natural selection encourage error-minimization? If selection for an error minimizing genetic code predated immutability and universality, then why is the standard code less than optimal?

Numerous hypotheses have been proposed to reconcile this apparent paradox [3][4][5][6]. It has been hypothesized that neutral evolution, for instance through proto-tRNA duplication (also termed "expansion"), could account for the code's near optimality (though not necessarily its universality) without the need for selection [6][7]. Other models have suggested that the code's progression might be explained entirely by selection for the best combination of genetic code and genome in a greedy fashion; however, these models are prone to premature freezing, particularly if the genome evolves rapidly [5][8]. Here we introduce an evolutionary model based on information-asymmetric games, which allow for a rich combination of both neutral evolution and selection, leading in combination to the suboptimal yet stable genetic code described above. The rest of the paper is organized as follows: We begin with a review of information-asymmetric games in the context of various applications of game theory to biology. We then describe a novel application of information-asymmetric games to the evolution of the genetic code. We compare our model's results to those from other competing models. Finally we conclude with a discussion of several implications of our model to the evolution in the RNA world.

As suggested by Maynard-Smith [9], games in a biological setting, unlike traditional ones in game theory, might not require "rational agents." A population of animals of the same species, for instance, may over the course of evolution behave according to game-theoretic principles even though none of those animals is a "rational agent," in a traditional sense. A species may "learn" over evolutionary time to select certain behaviors through random mutations, genetic drift, and selection, and ultimately reach a Nash equilibrium, in this case defined as an evolutionarily stable state in which each agent does not deviate strategies so long as all other agents in the system also do not deviate from their adopted strategies. "Utility" in the game-theoretic sense physically manifests as reproductive fitness. It is also common in nature that the interactions between such players (i.e. organisms of different species) will be asymmetric, due to, for example, differences in size or speed. The properties of such games have been studied extensively [9]. The asymmetry is often key to the game's equilibrium, for example when certain agents may learn to retreat when facing a member of a more dominant species.

An information-asymmetric (or "signaling") game is a particular kind of asymmetric game in which the asymmetry is defined by a difference in information each agent has about the state

of the system. Agents with information are termed "senders," and those without are termed "receivers." Receivers cannot observe the information the senders have directly; however, they can act according to "signals" or messages observed from the sender. Such a game may have many possible equilibria [10]. Senders and receivers may evolve a signaling convention ("separating" equilibrium) in which the sender sends a signal, correlated with the state of the system, to the receiver, whose actions are in turn correlated with signals observed. In this way information is passed via a signaling convention from the sender to receiver. It is also possible that the sender will send messages that are random with respect to the state of the system, or that the receiver will perform a random action or the same action regardless of the signal sent by the sender; in these cases the senders and receivers are in an uncoordinated equilibrium (for a more complete review of signaling games see [11]).

There are many instances in biology where signaling games provide a suitable abstract framework to describe and reason about how a set of agents might coordinate to overcome an inherent information asymmetry. In one well-studied system, agents function as both senders and receivers in a prisoner's dilemma-like game. Agents might have access to an arbitrary signal that is initially uncorrelated to strategy but becomes correlated over the course of evolution (dubbed the "green-beard" effect) [12][13]. Many molecular processes, from traditional "signaling" pathways to the translation of DNA/RNA to proteins, can be described using signaling games, though often senders and receivers must be treated as separate agents. In molecular systems, agents' "behaviors" are the chemical attributes of the senders or receivers, i.e. what molecules they react with and how they react (via conformational changes or the formation and breakage of chemical bonds) [14]. In the case of the genetic code, we envision a game between proto-mRNA (strings of codons with information) and set of proto-tRNA (RNAs with distinct anticodons, each able to bind a particular amino acid). Importantly, although proto-mRNA has information, it is unable to act (synthesize peptides) and proto-tRNA, though it is able to act, must rely on proto-mRNA for information regarding what constitutes a useful ordering of amino acids.

In signaling games, a utility function maximizing information flow between sender and receiver leads to stabilization of a separating equilibrium. If there exist many possible signaling conventions, as would be the case if one were to compare hypothetical alternate genetic codes, the conventions maximizing information transfer between senders and receivers would be favored. However, we must also consider that messages sent by the senders may be complex, consisting of a chain of more elementary signals. Longer messages might yield greater utility, but only if the message as a whole is transmitted and acted upon correctly. For example, it is possible that an enzyme with 100 amino acids is able to act with greater catalytic effectiveness than one with 10 amino acids, but only if the longer peptide is translated accurately.

Thus we conceive a framework in which the utility of agents is proportional to the message length (rate) and inversely proportional to the error in information transfer (distortion). The use of rate-distortion has previously been used to describe the effectiveness of protein translation in the context of information theory [15][16]. At this point, one might expect

that senders will send the longest messages ("proteomes") possible and that senders and receivers will select a signaling convention minimizing distortion. However, there is an additional constraint on the potential for tRNA to mutate: in a system in which longer messages have been established around a signaling convention, the cost of experimenting with new signaling conventions increases. For example, in the modern genetic code the nucleic acid sequence CUG codes for the amino acid Leucine. If a mutation in tRNA or aminoacylation machinery were to mutate so that CUG now codes for Serine, the progeny would have a higher likelihood of being viable if there were fewer CUG codons throughout the proteome[1].

Extended Signaling Evolution Framework

It is usually hypothesized that the genetic code formed in the context of an RNA world, gradually exposed to an emerging amino acid world [3][4]. We envision a scenario with two agents: proto-mRNA (strings of codons with information) and sets of proto-tRNA (RNAs with distinct anticodons, each able to bind a particular amino acid). In a given generation proto-mRNA and a particular set of proto-tRNA interact. The pair replicates via RNA-replicase ribozymes. However, they may also chemically aid their own replication through the accurate production of proteins (possible identities of these proteins are stipulated in Discussion).

We consider two cases: one in which these two agents may be co-localized in protocells, in which case they would evolve together and share a mutual utility function; and one in which they are separate entities in a syncitia-like setting, in which case mRNA and tRNA evolve separately by shuffling in every generation [5]. Utility for the cell (or interaction) is proportional to both message length (proteome encoded by the mRNA) and the probability the entire proteome is translated correctly, which depends on the robustness of the genetic code as encoded by the tRNA set. Error-minimization of the genetic code is especially important for supporting longer proteomes. For example, a false amino acid incorporation rate of 0.010 per codon per translation would allow a 10-amino acid polypeptide to be translated correctly 90% of the time and a 100-amino acid polypeptide 37% of the time. However, a false amino acid incorporation rate of 0.005 per codon per translation would translate a 100-amino acid polypeptide correctly 61% of the time.

During the replication process, each mRNA and tRNA may mutate with small probability, acquiring new message lengths via gene duplication/deletion and new genetic codes with different error rates via mutation in proto-tRNA anticodons. The probability that proto-tRNA will mutate decreases exponentially with proteome length of mRNA for aforementioned reasons. Equilibrium is reached when species with only one genetic code dominate.

---

[1] In the vast majority of organisms the genetic code is universal and immutable. However, in mitochondria, with relatively short proteomes, reassignments have been observed and occur with greater frequency as proteome length decreases [17].

We construct a framework for simulating the dynamics of the extended signaling games described above (see Mathematical Description of Framework). Although the parameters used in the simulations presented here are inspired by the specifics of genetic code evolution, the framework could be applied to any such game where message length and distortion for a given code are variable. In this system there are mRNA with many possible message lengths (N; in this case 10, 20, … 100 signals) and with genetic codes of different error rates (E; in this case 0.05, 0.045, … 0.005 errors per signal per generation), although we purposefully leave the structure of those codes unspecified.

We first consider the case where organisms are colocalized in protocells (Figure 1). Because of the colocalization, the mutation and evolution of tRNA and mRNA are intertwined. By defining a "species" as a colocalized pair of mRNA and tRNA, we can thus represent all possible species (as well as their potential mutated progeny) by the species map in Figure 1. The proto-mRNA 10-mers can mutate (through duplication) to produce multimers of greater utility. In addition, the genetic code can acquire "error-minimizing" reassignments (i.e., ones in which errors do not unduly penalize the system). The simulation begins with all organisms existing in the state of an uncoordinated equilibrium, but at a certain time (t=0) one organism spontaneously acquires the ability to encode a 10-mer using a genetic code with error rate E=0.05 (it is more likely for organisms to escape non-separating equilibria by first encoding short messages with a code that is not necessarily error-tolerant; see Discussion for details). The evolution of the population is modeled using ordinary differential equations, which approximate a large, well-mixed population, as well as simulations allowing stochastic events such as extinction in a discrete population of limited size. Population pressure simulating competition (for ATP or other nucleotides) is implicit in a death rate that is proportional to the total population number.

Next we consider the case where mRNA and tRNA are not colocalized and are instead part of a larger syncytium. In this case, proto-mRNA and proto-tRNA can still be described by the states described above, but their fates are not intertwined beyond a single generation. In addition, in a model of evolution of such separate agents, signaling game theory would suggest that, because of the information asymmetry, after a signaling convention is established, it might be possible for senders to "deceive" receivers into acting in a way that benefits senders but not receivers[2] [11]. In a situation where utility is shared equally between senders and receivers, as is the case in co-localization in cells, such deception is not beneficial to the deceiver. However, we explore this possibility in the model by introducing a third type of agent: "deceptive" mRNA, which encodes proteins beneficial to the reproduction of the "deceptive" mRNA but not proto-tRNA that does the actual translation. We present the results of simulating a syncytial population with and without deceptive mRNA.

[2] In the "green-beard altruism" game discussed in the introduction, it has been shown that deception enters into the game continuously, forcing players to cycle through signaling conventions rather than stay in a single separating equilibrium [12][13].

Results

As shown in Figure 1, in a simulation of a protocellular population in which organism abundances are discrete, it is possible for a species with a suboptimal signaling convention to dominate the population in a stable manner. Thus, in these simulations, exploration of alternate signaling conventions causes organisms in a population to concentrate toward error-minimal signaling conventions. Similarly, organisms concentrate toward longer messages, but only when the selected signaling convention is robust enough to support them. However, once species of longer message length dominate the population, the adopted signaling convention becomes immutable because the cost of deviating from the accepted convention rises steeply. The population's adopted convention thus becomes "frozen" in a suboptimal state.

We note that a model with small population size and discrete organisms is important; in a model based on ODEs resembling a large, well mixed population, some organisms would appear "clairvoyant" and would always reach the most optimal genetic code. We also compare our results with those of previous models of genetic code evolution [8] based on greedy selection of codes and genomes. While the optimality of genetic code evolution in those "myopic" models is governed solely by the probabilities that a new genome or genetic code will be "encountered," our model, which simulates actual competition between organisms of different fitness, demonstrates that the genetic code must reach a state that is "optimizing enough" to support a longer proteome before it can dominate. Thus our model predicts that there is a "sufficient" error rate for which the genetic code will freeze (Figure 2: mRNA and tRNA in protocells).

Others have hypothesized that heavy intercellular communication, with rapid fusion and division of cells [5] might have been common and would encourage universality by allowing more efficient codes and genes to selectively sweep through a "syncytial" population. A model of mRNA and tRNA in a syncytium without the possibility of deception does predict a higher likelihood of reaching an optimal genetic code because more optimal tRNA sets, which are likely to emerge via mutation when paired with short mRNAs, are free to associate with longer mRNAs in subsequent generations. However, if deceptive mRNA is a possibility, once tRNA evolves a robust genetic code, the deceptive mRNA will proliferate rapidly, leading to extinction of the both non-deceptive mRNA and tRNA (Figure 2: Separate mRNA and tRNA "syncytia"). Thus, a model based on signaling game theory would predict that genetic code optimization could have, in fact, encouraged the emergence of cellularity. Such an arrangement forces a shared utility function between mRNA and tRNA, thus averting deception.

Discussion

We have shown that a near-optimal, near-universal, and immutable code is a reasonable outcome from an information-asymmetric game in which utility is related to both message

length and error minimization. The model presented here demonstrates that the modern genetic code evolved most likely by a combination of previously hypothesized forces, involving neutral and selective evolution. Whereas a natural predisposition toward an error-minimizing code is not a necessary condition for an optimized genetic code, neutral evolution may have been an important force in establishing universality. At the same time, selective pressure can provide a powerful impetus for a genetic code to move toward error-minimization and, somewhat surprisingly, also enforce its immutability so as to maintain compatibility with the genome.

The formalization presented here in terms of a signaling game between proto-mRNA sender and proto-tRNA receiver assumes an RNA world becoming exposed to an emerging amino acid world. Evidence for such a world, for example, in the form of self-aminoacylating RNA [18][19], is growing. However, associated with such a world are a number of complexities, which we have ignored here. For instance, we have not accounted for the possibility that such an amino acid world would have transformed as a result of, for instance, biosynthesis of novel amino acids [4]. It is also possible that there were significant RNA machinery involved in translation itself, which were later replaced by peptides. The introduction of new amino acids and ribozymes and the extinction of others would undoubtedly affect genetic code evolution and could disrupt any signaling equilibria established by proto-mRNA and proto-tRNA.

Our mathematical framework could have also incorporated additional restricting factors, such as the natural affinity of certain tRNAs for certain amino acids due to stereochemical constraints [3]. The genetic code also has evolutionarily beneficial properties [20], which may have allowed for additional genome regulation or preservation of RNA sequences with catalytic activities. It is also possible that nucleotide and amino acid concentrations were non-uniform in primordial conditions, tilting the structure of the genetic code so that codons composed of abundant nucleic acids coded for the most necessary or most common amino acids. Furthermore, certain codons-anticodon pairs are more thermodynamically stable than other pairs. These pairings likely influenced the development of the genetic code; for example genes that are efficiently expressed take advantage of the bias by over-utilizing codons with strong codon-anticodon interactions [21]. Such constraints may lead to likely flow patterns between the different genetic code states, akin to “natural salience” for certain words in an evolving language [11].

In a similar vein, one may question how the first RNA-based proto-cells could have escaped a non-separating equilibrium when first exposed to amino acids. Because of the relative length and complexity of modern enzymes, it may be possible that the earliest peptides were not enzymes in the traditional sense. To “accidentally” stumble upon genes encoding such enzymes at the same time an error minimizing code occurred by chance, as suggested by Crick [3], has vanishingly small probability. Our signaling model suggests that there existed a reproductive benefit to encoding even short strings of polypeptides while the usage of each codon was still nascent. Even short proteins might have provided structural support to proto cells or had catalytic activity with specificity, as has been observed in some dipeptides [22].

Incorporating amino acids into an elongating polypeptide using a nucleic acid template may also have prevented those amino acids from doing harm to cells. Experiments have shown that proteins of random sequence may have exposed hydrophobic surfaces, which aggregate [23]. Homopolymers behave in a similar manner; their accumulation can be lethal to cells [24]. A more accurate genetic code, if nothing else, would have allowed a proto-cell to package amino acids into soluble globules. One could further envision a path by which soluble proteins lead to proteins with biological functions, as has been observed in *in vitro* evolution experiments [25][26][27].

Overall this paper presents a framework for studying signaling game dynamics in instances where both message length and distortion are factors in the utility of both senders and receivers. Although we have applied the framework here primarily to the evolution of the genetic code, similar analyses might be applied to the evolution of many other seemingly fixed processes [28][29], where the evolutionary clock appears to have frozen a biological process prematurely to an arbitrary conventional structure.

Mathematical Description of Framework

Population dynamics

We construct a population model inspired by the dynamics described in [30][3]. In this model, organisms can be one of several species of a given signaling convention (genetic code) and message (proteome) length. The change in organism number of a species S_i is given by

$$\frac{dS_i}{dt} = S_i\left(b\left(1 - \sum_{j\in\mathcal{L}} \mu_{length:i\to j} - \sum_{j\in\mathcal{G}} \mu_{c:i\to j}\right) - d\right) + \sum_{j\in\mathcal{G}} S_j \mu_{c:j\to i} + \sum_{j\in\mathcal{L}} S_j \mu_{length:j\to i}$$

(1)

where $\mathcal{L}$ is the set of species with the same genetic code as species i with lengths attainable by one mutation (neighbor lengths); $\mathcal{G}$ is the set of species with the same length as species i with genetic codes attainable by one genetic code mutation (neighbor genetic codes). Although the notation here implies a set of ordinary differential equations, we use the same construct as the basis for stochastic simulations, with discrete population levels, in which b and d are probabilities that an organism will reproduce or die in any generation, rather than an average

[3] Note that our mathematical model bears certain similarities to Eigen's model of hypercycles [31], which is described by similar ODEs and addresses the emergence of complex interactions between species in an RNA world. Eigen's hypercycles describe self-reproducing molecular systems, in which RNAs and enzymes "cooperate" to enable the enzymes to cyclically increase RNA's replication rates. Our model provides a direct game-theoretic framework to interpret emergence and stability of such cooperation.

rate. In the syncytial model considered in Figure 2D,E all genetic codes and proteomes are randomly shuffled between generations. When deceptive mRNA is introduced to the population, it also pairs with tRNA, but when it does only the deceptive mRNA is replicated. We expand on the birth rate b, rates of mutation μ_{length} and μ_c, and the death rate d below.

Species Fitness.

We ascribe an organism's reproductive success in part to the probability that the organism will be able to correctly translate its proteome throughout its life cycle. We define $p_{correct}$, the probability that an organism will be able to synthesize its proteome without errors each full transcript translation event, as

$$p_{correct} = \prod_{i \in A} \prod_{j \in A} (1 - p_{j \to i})^{k_{j \to i} n_{proteome:\, j}}$$

(2)

where A denotes amino acid space, and $p_{j \to i}$ denotes the probability that amino acid j would be replaced by amino acid i during translation. Note that the quantity $p_{j \to i}$ is equivalent to the error rate per codon per translation/replication event (E in Figure 2). The parameter $k_{j \to i}$ denotes the physiochemical similarity between the amino acids i and j. For maximally dissimilar amino acids, i and j, the parameter $k_{j \to i}$ =1. This parameter could be assigned based on a physiochemical distance matrix such as the Grantham matrix [32]. The parameter $n_{proteome:\, j}$ denotes the number of j amino acids in the proteome. Note that mutations could affect an organism's reproductive success in a similar manner.

Note that tRNA misincorporation takes place on the order of 10^{-3} per codon per translation in yeast [33], and mutation takes place on the order of 10^{-3} per base pair per replication in RNA viruses [34]. We assume the rate of mutation or misincorporation would be higher in the context of an RNA world.

We postulate there is a benefit to having longer proteomes, allowing for greater biochemical complexity. For simplicity, we represent this benefit as a linear relationship between reproductive rate and the length of the portion of the genome that encodes proteins. Thus the fitness of a species with a genetic code resulting in a certain $p_{correct}$ upon reproduction and a genome of length $n_{genome:\, tot}$ is given by

$$b = b_0 p_{correct} n_{genome:\, tot}$$

(3)

where b_0 is a constant indicating the number of generations per unit time, of an organism of maximal genome length and error-free translation. Note that $n_{genome:\,tot} = 0$ would correspond to a situation in which proteins encoded by the genome are dysfunctional.

Mutation to different species

A certain percentage of the progeny born to a given species will be mutants, either acquiring genomes of different length or acquiring a different genetic code. Note that we ignore other types of mutation not affecting genome length or genetic code composition. The relative chance (as a percentage) that a genome will acquire a different length through insertions, gene duplication events, deletions, etc. is a constant, μ_{length}. Note that the changes in length could conceivably be any integer greater than or equal to 1, depending on the mechanism of length modification.

Mutation in the genetic code may be more difficult as codons become assigned and used [17]. If multiple triplets in the genetic code encode the same amino acid, we assume genetic drift will allow those triplets to interchange. Thus the probability that one of those triplets will not be used at all, allowing it to be reassigned to a new amino acid, can be determined using an extension of Wright-Fisher model for allele frequency at equilibrium [30][35]. The probability of a genetic code change in which codon *c* codes for amino acid *x* in the mother organism, and codes for a different amino acid *y* in the daughter organism, is given by

$$\mu_{c:y\to x} = \mu_0\left(\frac{n_{GC:\,x} - 1}{n_{GC:\,x}}\right)^{k_{y\to x} n_{genome:\,x}}$$

(4)

where: μ_0 is the rate of mutation of tRNA, allowing for either aminoacylation of a codon by a new amino acid, or mutation of an anticodon loop in a copy of a tRNA corresponding to a different codon; $n_{GC:\,x}$ is the number of codons in the genetic code which encode amino acid *x*; and $n_{genome:\,x}$ is the number of times amino acid *x* is coded for in the genome. Note if multiple copies of a tRNA exist, ambiguity in codon assignment could be resolved only through the elimination of either the old or new isoform [36]. We also incorporate the physiochemical similarity parameter $k_{y\to x}$ as described in equation (1) because it may be the case that a mutation resulting in more similar amino acids may be less disruptive to function than a mutation to highly dissimilar amino acids. If certain genes can accept either amino acid *x* or amino acid *y*, $n_{genome:x}$ decreases accordingly[4].

[4] Note that these assumptions also accord with the observation that in *Candida* the CUG codon was reassigned from leucine to serine, as leucine along with arginine has the greatest representation in code space. Although the *Candida* genome is long relative to mitochondria, there also is evidence that AT pressure could have acted to artificially increase the probability that the CUG codon is free [37].

Selective pressure

We model selective pressure on the population due to limited resources (ATP and other nucleotides) by imposing a death rate, which is proportional to the size of the population, given by

$$d = \frac{total\ population}{K}$$

(5)

where K is a constant carrying capacity. Negative selection is not explicitly modeled because for small mutation rates, with suitable re-parameterization, that model can be reduced to the one described here without affecting the observed results.

The parameter values used for the model in Figures 1 and 2 are: K=1000, $k_{y\to x}$=0.3, $n_{GC:\,x}$=2, b_0=1, μ_{length}=0.1, and μ_0=1. In Figure 2E, deceptive mRNA has a b_0=1.5, to simulate advantageous growth. Other values of K, μ_{length}, and μ_0 were also explored, along with conditions such as genetic code branching; for results, see Figures S1, S2, and S3.

We have also explored a model of fixed population size, which allows for multiple optimal genetic codes and explicit mutations, allowing one to trace the evolution of the genetic code in its entirety from pooling equilibrium to near-optimal, near-universal separating equilibrium (See SI).

Availability

The code used to implement the mathematics described in the main text and simulation described here is freely available at: http://bioinformatics.nyu.edu/projects/genetic-code/

Acknowledgment

JJ, AS, SEM, and BM designed the research. JJ performed the research. JJ and BM wrote the paper. This work was inspired by discussions with Heeralal Janwa, Michael Wigler, Andi Witzel, Loes Olde Loohuis, Rohit Parikh, and Alfredo Ferro. In particular, BM wishes to thank Profs. Janwa and Ferro for their invitation to Puerto Rico and Lipari, where many ideas from information and game theory were discussed. Prof. Wigler contributed many key biological ideas to the vision of the primordial world in this paper, such as the importance of peptide solubility.

Funding

This research was funded by two NSF grants: NSF CCF-0836649 & NSF CCF-0926166. JJ was supported by a National Defense Science and Engineering Graduate Fellowship from the US Department of Defense.

References

[1] Woese, C.R. (1965) Order in the genetic code. *PNAS*. 54:71-75.

[2] Alff-Sternberger, C. (1969) The genetic code and error transmission. *PNAS.* 64:584–591.

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

[4] Wong, J.T. (1975) A co-evolution theory of the genetic code. *PNAS*. 72(5):1909-12.

[5] Vetsigian, K., Woese, C., Goldenfeld, N. (2006) Collective evolution and the genetic code. *PNAS*. 103(28): 10696-10701.

[6] Osawa, S., Jukes, T.H. (1989) Codon reassignment (codon capture) in evolution. *J Mol Evol*. 28(4):271-8.

[7] Massey, S. (2008) A Neutral Origin for Error Minimization in the Genetic Code. J. Mol. Evol. 67(5):510-516.

[8] Sella, G. and Ardell, D.H. (2006) The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol*. 63(3):297-313.

[9] Mavnard Smith J & Parker G A. (1976) The logic of asymmetric contests. *Anim. Behav*. 24: 159-75.

[10] Cho, I-K, Kreps, D.M. (1987) Signaling Games and Stable Equlibria. *The Quarterly Journal of Economics*. 102(2):179-221.

[11] Skyrms, B. (2010) *Signals*. Oxford Scholarship Online. DOI: 10.1093/acprof:oso/9780199580828.001.0001

[12] Jansen, V.A.A., van Baalen, M. (2006) Altruism through beard chromodynamics *Nature* 440, 663–666,

[13] Traulsen, A., Nowak, M.A. (2007) Chromodynamics of Cooperation in Finite Populations. PLoS ONE 2(3).

[14] Smith, J.M. (1999) The Idea of Information in Biology. *The Quarterly Review of Biology.* 74(4):395-400.

[15] Tlusty, T. (2008) A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity, and cost. *Phys Biol*. 5 016001.

[16] Tlusty, T. (2010) A colorful origin for the genetic code: Information theory, statistical mechanics, and the emergence of molecular codes. *Phys Life Rev*. 100 048101.

[17] Massey, S.E. and Garey, J.R. (2007) A Comparative Genomics Analysis of Codon Reassignments Reveals a Link with Mitochondrial Proteome Size and a Mechanism of Genetic Code Change Via Suppressor tRNAs. *J Mol Evol.* 64(4):399-410

[18] Turk, R.M., Chumachenko, N.V., Yarus, M. (2010) Multiple translational products from a five-nucleotide ribozyme. *PNAS.* 107(10):2485-2589.

[19] Murakami, H., Ohta, A., Goto, Y., Sako, Y., Suga, H. (2006) Flexizyme as a versatile tRNA acylation catalyst and the application for translation. *Nuc Acids Symp Ser*. 50(1):35-6.

[20] Itzkovits, S., Alon, U. (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* 17(4):405-12.

[21] Grosjean, H., Fiers, W. (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18(3):199-209.

[22] Weber, A.L., Pizzarello, S. (2006) The peptide-catalyzed stereospecitic synthesis of tetrose: a possible model for prebiotic molecular evolution. *PNAS*. 103(34):12713-7.

[23] Mandecki, W. (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Engineering*. 3(3):221-226.

[24] Omo, Y., Kino, Y., Sasagawa, N., Ishiura, S. (2005) Comparative analysis of the cytotoxicity of homopolymeric amino acids. *Biochimica et Biophysica Acta Proteins & Proteomics.* 1748(2):174-179.

[25] Keefe, A.D., Szostak, J.W. (2001) Functional proteins from a random sequence library. *Nature*. 410(6829):715-718.

[26] Hayashi, Y., Sakata, H., Maniko, Y., Urabe, I., Yomo, T. (2003) Can an Arbitrary Sequence Evolve Towards Acquiring a Biological Function? *J Mol Evol*. 10.1007/s00239-002-2389-y.

[27] Ito, Y., Kawama, T., Urabe, I., Yomo, T. (2004) Evolution of an Arbitrary Sequence in Solubility. *J Mol Evol*. doi:10.1007/s00239-003-2542-2.

[28] Gerhart, J., Kirschner, M. (2007) The theory of facilitated variation. *PNAS* 104(Suppl1):8582-8589.

[29] Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo A., Dekel, E., Kavanagh, K., Alon, U. (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science*. 336(6085):1157-1160.

[30] Wright, S. (1931) Evolution in Mendelian Populations. *Genetics*. 16:97-153.

[31] M. Eigen (1971) Self organization of matter and the evolution of biological macromolecules, *Die Naturwissenchafen* 58, 467-523.

[32] Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science.* 185:862–864.

[33] Kramer, E.B. and Farabaugh, P.J. (2006) The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*. doi:10.1261/rna.294907

[34] Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., VandePol, S. (1982) Rapid Evolution of RNA Genomes. *Science*. 215:1577-85.

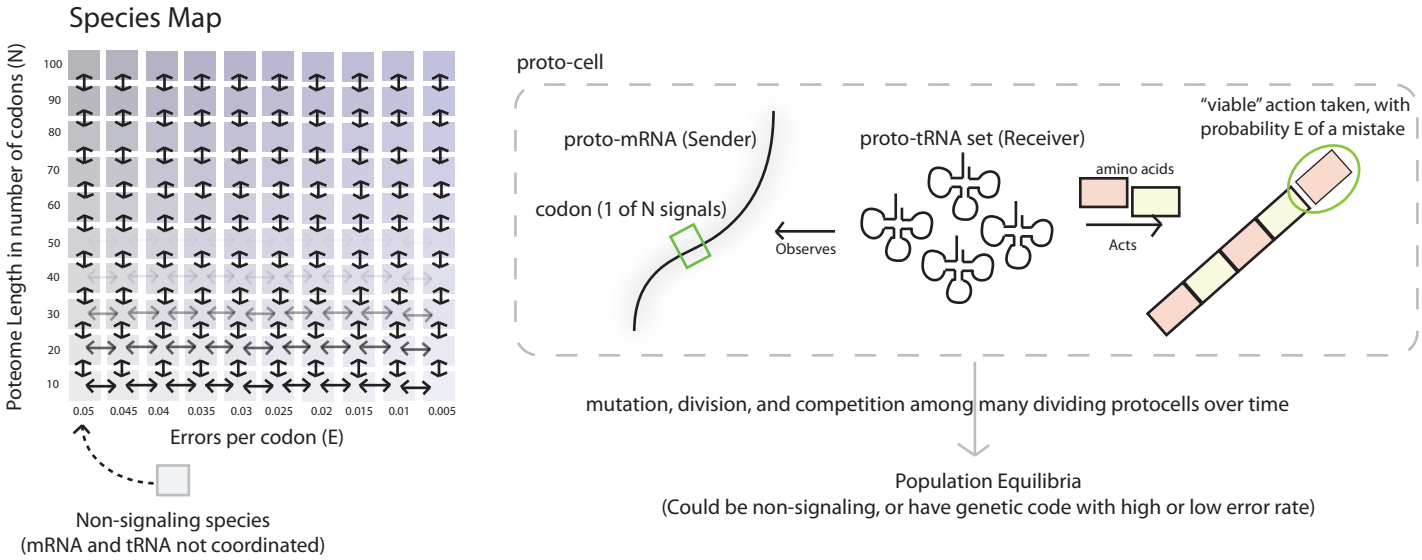
[35] Fisher, R.A. (1922) On the dominance ratio. *Proc. Roy, Soc., Edinb.* 42: 321-341.

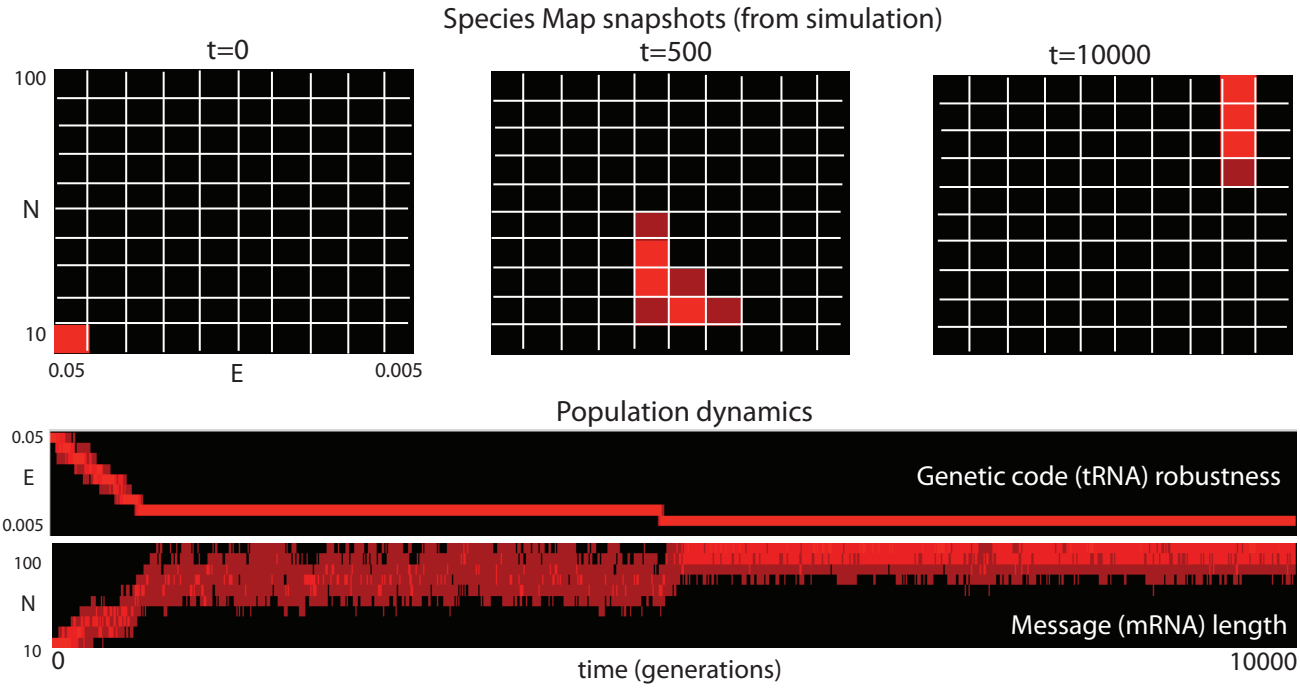
[36] Schultz, D.W. and Yarus, M. (1994) Transfer RNA Mutation and the Malleability of the Genetic Code. *J Mol Biol.* 235:1377-1380.

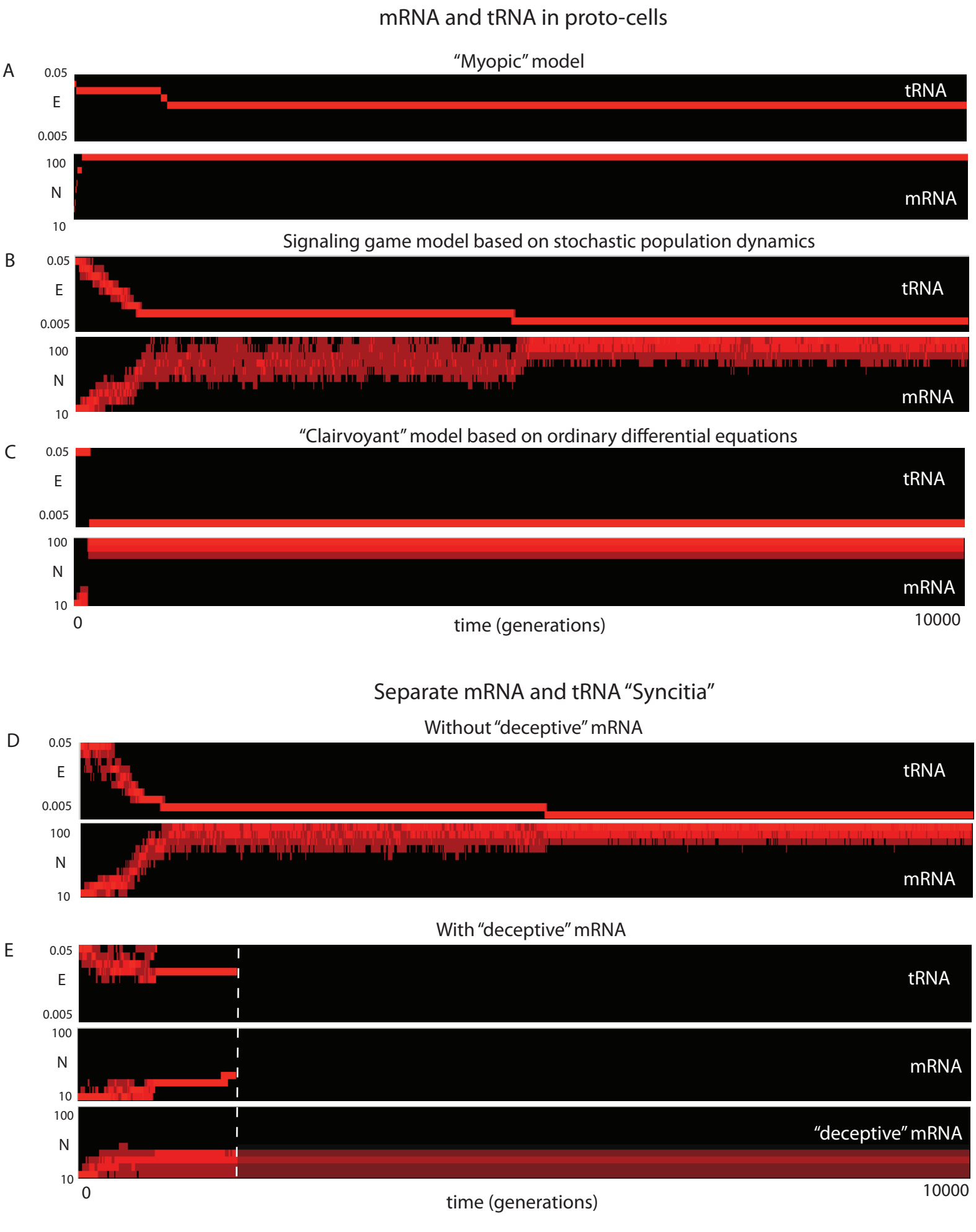
[37] Silva, R.M., Miranda, I., Moura, G., Santos M.A.S. (2004) Yeast as a model organism for studying the evolution of non-standard genetic codes. *Briefings in Functional Genomics and Proteomics*. 3(1):35-46.

Figure 1. *Species Map*. Depicted are the possible species of a given error rate per signal unit, E, and messages of length N. There are ten possible lengths (10, 20, … 100 signals) and ten possible codes with error rates E (0.05, 0.045, … 0.005 errors per signal per generation). During reproduction of a protocell of a given species, mutation is permitted between adjacent species on the map. Lighter arrows indicate lower mutation rates (see Mathematical Description of Framework). *Simulation Results*. Three snapshots of population levels at various time points are shown. In the first snapshot, the simulation begins with the birth of one organism of species E=0.05, N=10. In the second, several codes are competing for existence. By the third, the system has reached equilibrium. Below, the range of E and N values in the population is shown for all time points of the simulation. Around time T=5000 the population transitions from a stable genetic code to a more optimal one, but also becomes more "frozen" as a result of the ensuing elongation of average message lengths.

Figure 2. The range of E values (codes represented by tRNA) and N values (message lengths of mRNA) are reported across time from simulations given different assumptions. A) A greedy algorithm for genetic code and genome selection is used as described by other researchers [5][8]. As previously reported, these assumptions can lead to "myopic" premature freezing of the genetic code, particularly if the genome is highly mutable. B) Results from our stochastic game-theoretic simulation. C) Results from an ODE version of the game theoretic simulation. Here, because there is infinite population size, the system appears "clairvoyant;" there will always emerge one organism of the most optimal genetic code, which will dominate the system. D) We consider a syncytial model of evolution as in [5], except using a game-theoretic simulation rather than a greedy approach. We also find that a syncitia encourages further optimality of the genetic code. E) However, when the possibility of deception is introduced, deception by the sender can lead to extinction of all tRNA and non-deceptive mRNA (white dashed line). The faded out "deceptive" mRNA seen after the white line is not reproducing, as there is no tRNA to pair with.


Species Map
Poteome Length in number of codons (N)
100
90
80
70
60
50
40
30
20
10
0.05 0.045 0.04 0.035 0.03 0.025 0.02 0.015 0.01 0.005
Errors per codon (E)
Non-signaling species
(mRNA and tRNA not coordinated)
proto-cell
proto-mRNA (Sender)
codon (1 of N signals)
Observes
proto-tRNA set (Receiver)
amino acids
Acts
"viable" action taken, with
probability E of a mistake
mutation, division, and competition among many dividing protocells over time
Population Equilibria
(Could be non-signaling, or have genetic code with high or low error rate)


Species Map snapshots (from simulation)
t=0
t=500
t=10000
100
N
10
0.05
E
0.005
Population dynamics
0.05
E
0.005
Genetic code (tRNA) robustness
100
N
10
Message (mRNA) length
0
time (generations)
10000


mRNA and tRNA in proto-cells
A
"Myopic" model
0.05
E
0.005
tRNA
100
N
10
mRNA
Signaling game model based on stochastic population dynamics
B
0.05
E
0.005
tRNA
100
N
10
mRNA
"Clairvoyant" model based on ordinary differential equations
C
0.05
E
0.005
tRNA
100
N
10
mRNA
0
time (generations)
10000
Separate mRNA and tRNA "Syncitia"
Without "deceptive" mRNA
D
0.05
E
0.005
tRNA
100
N
10
mRNA
With "deceptive" mRNA
E
0.05
E
0.005
tRNA
100
N
10
mRNA
100
N
10
"deceptive" mRNA
0
time (generations)
10000

Supporting Information

Moran Model Simulating a Primordial World

We construct a more concrete Moran model, a population simulation where the number of organisms is held constant throughout the simulation. As organisms reproduce, they randomly replace another organism in the population. The reproductive rate of the organisms is governed by their fitness (see below). The model we implemented begins with 100 organisms with random genome sequence and a genetic code of all stop codons. The genetic code of an organism is a circular array of length 9, where each location (codon sequence) in the array maps to either a red or green amino acid or stop codon. The genome is a linear array of up to 25 codons.

During each translation step, the genome is translated into a string of amino acids and stop codons via the genetic code. During this process, a codon may receive the amino acid encoded by one of its neighbors in code space with a probability of 0.05. At the end of the translation step, the polypeptide sequence is evaluated. The longest polypeptide of sequence {red-green-}^{N}-stop adds N points toward an organism's reproductive score. Other polypeptides are considered incidental and do not add to the reproductive score; however, during each translation step each organism receives 1 point toward its reproductive score because we assume its RNA has replicative function even in the absence of protein synthesis. Over multiple translation steps, each organism accumulates points toward its reproductive score. When it reaches a threshold of 100 points, the organism produces a daughter cell by replacing a random organism in the population. During replication, each codon in the genetic code encodes a different amino acid or stop codon with probability 0.01 per codon per replication, and each position in the genome mutates to a different codon with probability 0.01 per codon per replication. This process allows for the coevolution of the genetic code and the genome. Each trial of the simulation was run for 40,000 translation steps.

We note that another way in which our simulation differs from previous ones [2][3] is that the pressure to minimize genetic code entropy comes from competition between tRNAs during translation and not from mutation. This difference stems from a choice to consider error minimization from the perspective of robustness with regard to tRNA wobble and other thermodynamic effects rather than closeness in mutational space.

Results

We start from a simple pooling equilibrium, in which all codons are assigned to stop instructions, and genomes are random codon strings. During the course of the simulation, organisms are reproductively favored if they synthesize polypeptides of a certain pattern. We simplify the genetic code as a one-dimensional ring. During the translation process a signal for any particular codon in the ring may be misread as a signal for one of the neighboring

codons in code space. As in other studies [3], such a simplification allows us to view many of features of a real genetic code without specifying base pairs and positions, which may have varying substitution rates based on complex thermodynamic and copy number effects. During replication, a codon in the genome may mutate to any random codon at a rate of 0.01 per position per replication, and the genetic code may change one of its assignments to a random amino acid or stop codon with a probability of 0.01 per codon per replication. These rates are relatively high because we assume that prior to the advent of protein translation and replication machinery, the frequency of errors might have been higher.

We quantify the entropy of the genetic code based on the likelihood that an incorrect amino acid will be incorporated during the translation process (Figure S4A: Genetic Code Assignments). After 100 trials, the genetic codes of organisms at the end of the simulation have low entropy when compared to random codes (Figure S4B,C). For illustrative purposes, the dominant lineage from one simulation is shown in Figure S5A. In this example the genetic code goes through a diversifying step followed by a consolidation step, at the end of which neighbor codons tend to code similar amino acids, as expected [1][3][4][5]. The consolidation step is the result of a single common ancestor emerging as the dominant organism in the population, although its spread can be attributed to either drift or fitness advantage. Thus, during the diversification step, there is also a wider degree of heterogeneity in the codes of the population as a whole, in agreement with the stochastic simulations presented in the main text. In addition, there is a wider degree of genomic heterogeneity during the diversification step, as the variety in genetic codes allows for greater experimentation with different genomes.

Further reduction in entropy is afforded by codon usage bias in the genome, as expected from observations of real genomes [6]. In our simulations, the equal and concentrated presence of both red and green amino acids manifests as a large and equal representation of those amino acids in code space. By contrast, the stop codon, which is used only once during the translation of a long peptide, tends to assume a smaller portion of code space.

In this setting, proteome length increases when codons abutting an existing gene mutate so that they encode amino acids in a proper sequence, which lengthens the protein-encoding gene (See Figure S4A: Proteome). A plot of the length of the longest encoded protein as a function of the entropy of the genetic code is shown in Figure S4C. As expected, longer genes do not emerge if the genetic code is not accurate enough to translate them properly. As shown by the red points in Figure S5C, dominant species at the end of the simulation tend to have long genes and genetic codes that are error-minimizing when compared to random codes. The range of entropy values observed reflect the tension between the need to encode a diversity of amino acids and the need to reduce entropy; this finding accords with previous theoretical studies [4][5].

In 66% of 100 trials, 80% or more of the organisms at the end of the simulation used only one genetic code. In the vast majority of these cases, the organisms not using the dominant code were related to a common ancestor using the same code and varied by one mutation.

Universality of the genetic code was established at some point in time in virtually every trial; however, in many cases after universality was established other derivative genetic codes emerged, most likely a result of the fact that genome length in our simulations was bounded to be at most 25 codons. Thus a simpler theory without a need to enforce horizontal gene transfer may suffice to explain universality.

Availability

The code used to implement the mathematics described in the main text and simulation described here is freely available at: http://bioinformatics.nyu.edu/projects/genetic-code/

References

[1] Vetsigian, K., Woese, C., Goldenfeld, N. (2006) Collective evolution and the genetic code. *PNAS*. 103(28): 10696-10701.

[2] Ardell DH, Sella G (2002) No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Phil Trans R Soc Lond B* 357:1625–1642

[3] Sella, G. and Ardell, D.H. (2006) The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol*. 63(3):297-313.

[4] Tlusty, T. (2008) A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity, and cost. *Phys Biol*. 5 016001.

[5] Tlusty, T. (2010) A colorful origin for the genetic code: Information theory, statistical mechanics, and the emergence of molecular codes. *Phys Life Rev*. 100 048101.

[6] Moriyama, E.N., Powell, J.R. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Eschericihia coli*. (1998) *Nucl Acids Res* 26(13):3188-3193.

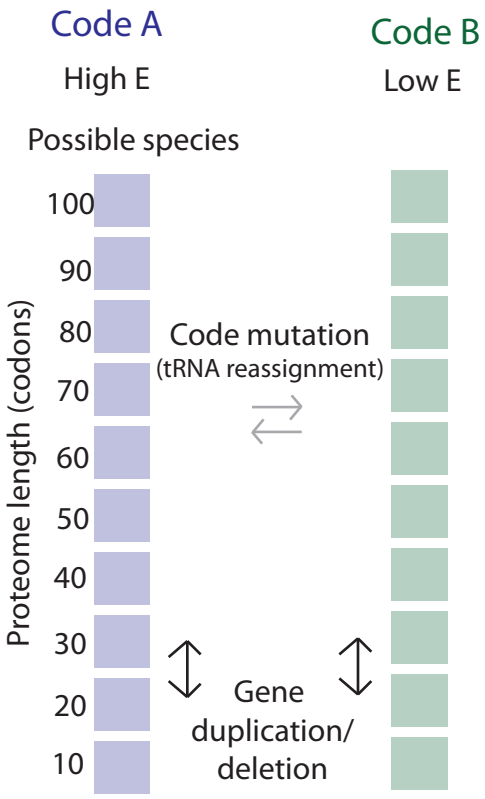
Figure S1. Simulations with two possible conventional signaling equilibria with different error tolerance. (A) Here, species may have one of two possible genetic codes, A and B, and possible proteome lengths 10, 20... 100. Code B (E=0.010) is error minimizing relative to Code A (E=0.015). It is possible to mutate from code B to code A (and vice versa) with a point mutation in a single tRNA. (B) The number of organisms of a given genetic code (A or B) is plotted against time, starting with one organism of A, proteome length=10. The plot titled "Normal" uses the same parameters as in Figure 2. Similarly, "Error(A) = Error(B)" depicts a simulation in which Code B has the same E value as Code A. "Low code mutation rate" depicts a simulation in which μ_0=0.1. "High code mutation rate" depicts a simulation in which μ_0 =4. "Low genome length mutation rate" depicts a simulation in which μ_{length} =0.01. "High genome length mutation rate" depicts a simulation in which μ_{length}=0.5.

Figure S2. A species map depicts three genetic codes. Code B and B* (E=0.010) are error minimizing relative to Code A (E=0.015). It is possible to mutate from code B to code A (and vice versa) and from B* to A (and vice versa) with a single tRNA mutation. Organisms with both codes may acquire genomes of longer length according to the dynamics previously described. In the six subplots shown, the number of organisms of a given genetic code (A, B or B*) is plotted against time, using the same parameters described in Figure S1.

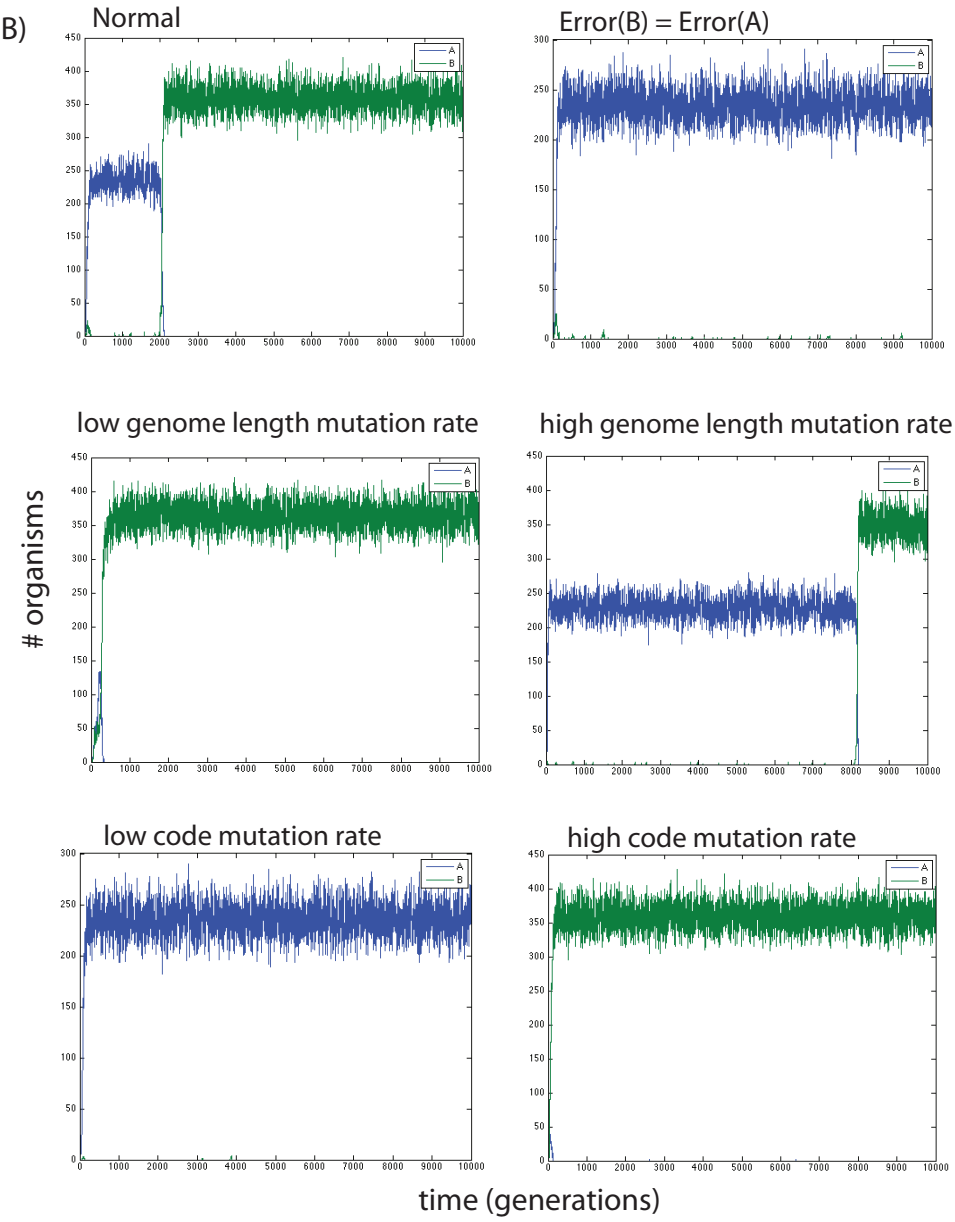
Figure S3. Stochastic simulations run according to the same setup as the simulation in Figure 1, but with varying carrying capacities, K. In the top row, the genetic codes present in the population at a given time (as represented by their error value E) are shown across time for three different simulations, K=100, K=1000 as in Figure 1, and K=10000. In the bottom row, a snapshot of the population levels at the end of each simulation is shown according to the species map in Figure 1.

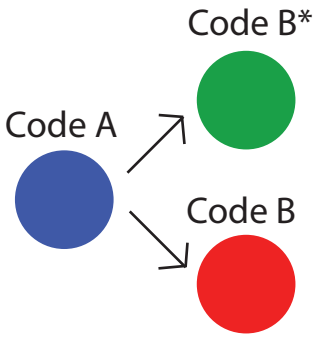
Figure S4. A) Schematic of a protocellular environment in which self-reproducing RNA vesicles containing both proto-mRNA (Genome) and proto-tRNA (Genetic Code) are immersed in a pool of highly concentrated amino acids (See SI: Primordial World Simulation for details). Genetic Code Assignments. The codon-amino acid assignments for one lineage are represented by a one-dimensional array. Any codon may substitute for its neighbors in code space with 5% probability. The genetic code assignments across generational time trend toward clustering similar amino acid assignments in blocks to reduce entropy, as well as reducing the representation of stop codons, which are used infrequently. To understand the evolution of proteome complexity, the longest protein encoded by the genome given the genetic code assignments above is recorded across generational time. B) The entropy for 10,000 randomly assigned genetic codes is shown for comparison with error-minimizing codes, which evolved in (C). C) Dominant species' genetic code entropy and encoded protein length from the end of 100 simulations are shown in red. Similar points from all ancestors of those final species are shown in blue.

A)

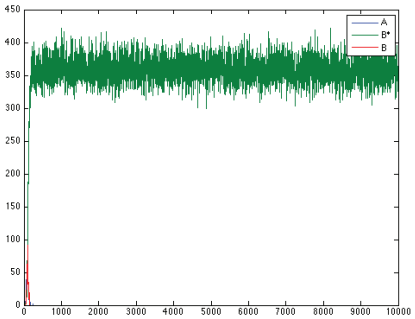

Code A
High E
Code B
Low E
Possible species
Proteome length (codons)
100
90
80
70
60
50
40
30
20
10
Code mutation
(tRNA reassignment)
Gene
duplication/
deletion

B)

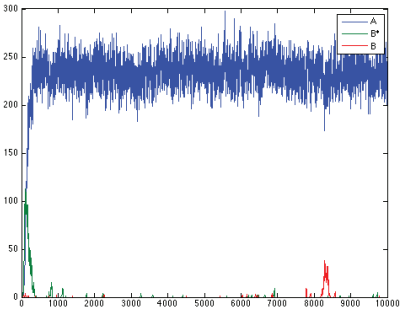

Normal
Error(B) = Error(A)
low genome length mutation rate
high genome length mutation rate
low code mutation rate
high code mutation rate
organisms
time (generations)
A
B



Normal

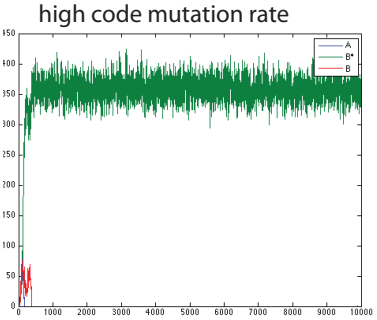


E(B) = E(B*) = E(A)

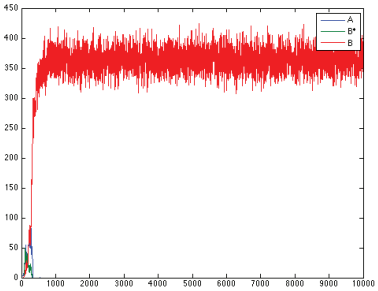


low code mutation rate

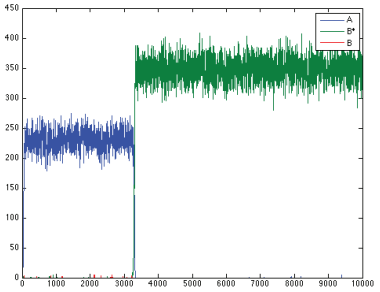
high code mutation rate



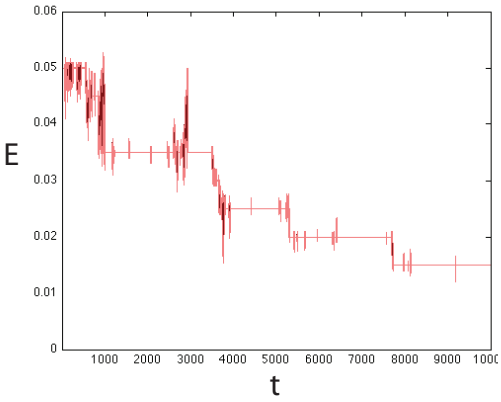
low genome length mutation rate



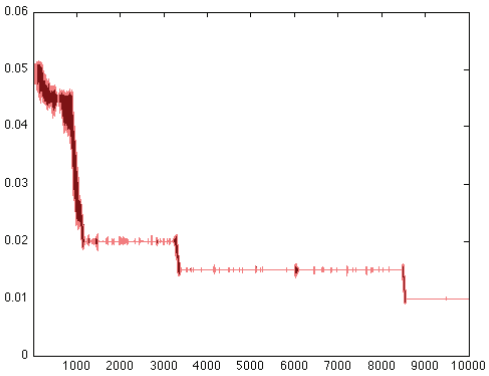
high genome length mutation rate



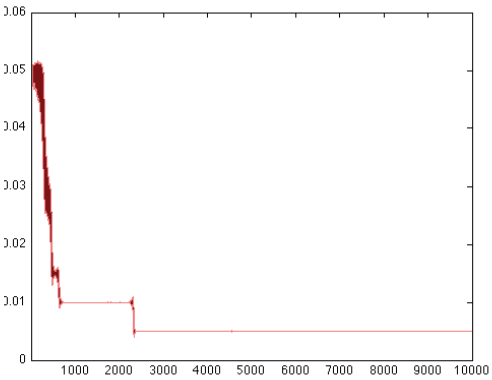
K = 100

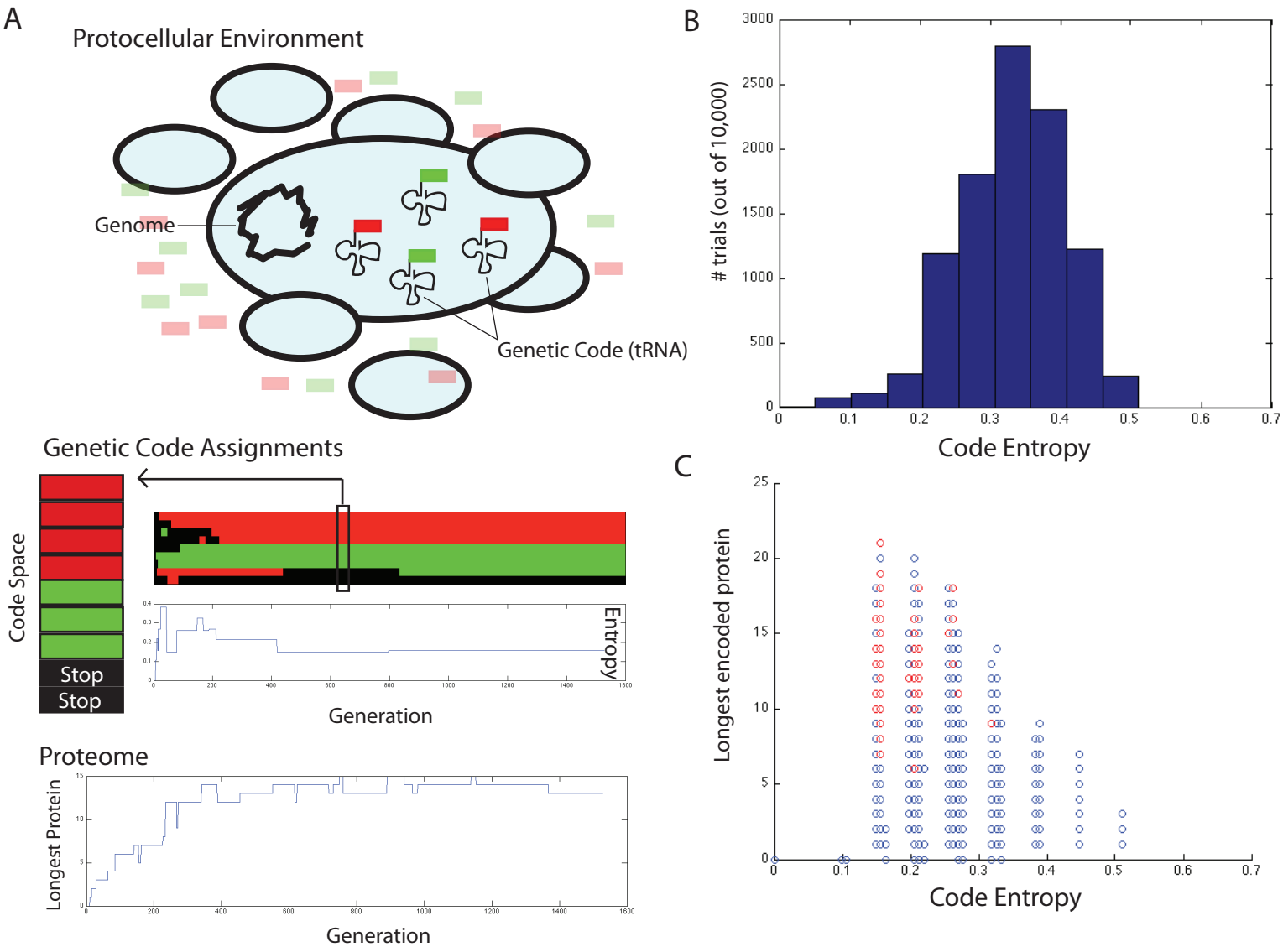


K = 1000



K = 10000




A
Protocellular Environment
Genome
Genetic Code (tRNA)
Genetic Code Assignments
Code Space
Stop
Stop
Entropy
Generation
Proteome
Longest Protein
Generation
B
trials (out of 10,000)
Code Entropy
C
Longest encoded protein
Code Entropy