JOURNAL OF THE ROYAL SOCIETY
Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Jee J, Sundstrom A, Massey SE, Mishra B. 2013 What can information-asymmetric games tell us about the context of Crick's 'frozen accident'? J R Soc Interface 10: 20130614.
http://dx.doi.org/10.1098/rsif.2013.0614

Received: 9 July 2013
Accepted: 7 August 2013

Subject Areas:
biomathematics

Keywords:
information-asymmetric games, signalling, genetic code, RNA world

Authors for correspondence:
Justin Jee
e-mail: justin.jee@med.nyu.edu
Bud Mishra
e-mail: mishra@nyu.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2013.0614 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

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] and Bud Mishra[1,2]

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

This paper describes a novel application of information-asymmetric (signalling) 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 signalling 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 near-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.

1. Introduction

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 [3–6]. 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 manner; 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 with 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 & Parker [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 behaviours 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 'signalling') 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 signalling 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 signalling 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 signalling games, see [11]).

There are many instances in biology where signalling 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 'signalling' pathways to the translation of DNA/RNA to proteins, can be described using signalling games, though often senders and receivers must be treated as separate agents. In molecular systems, agents' 'behaviours' 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 sets 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 signalling games, a utility function maximizing information flow between sender and receiver leads to stabilization of a separating equilibrium. If there exist many possible signalling conventions, as would be the case if one were to compare hypothetical alternative genetic codes, then the conventions maximizing information transfer between senders and receivers would be favoured. 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) but is restricted by 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 signalling 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 signalling convention, the cost of experimenting with new signalling 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]

2. Extended signalling 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 colocalized 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 syncytia-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

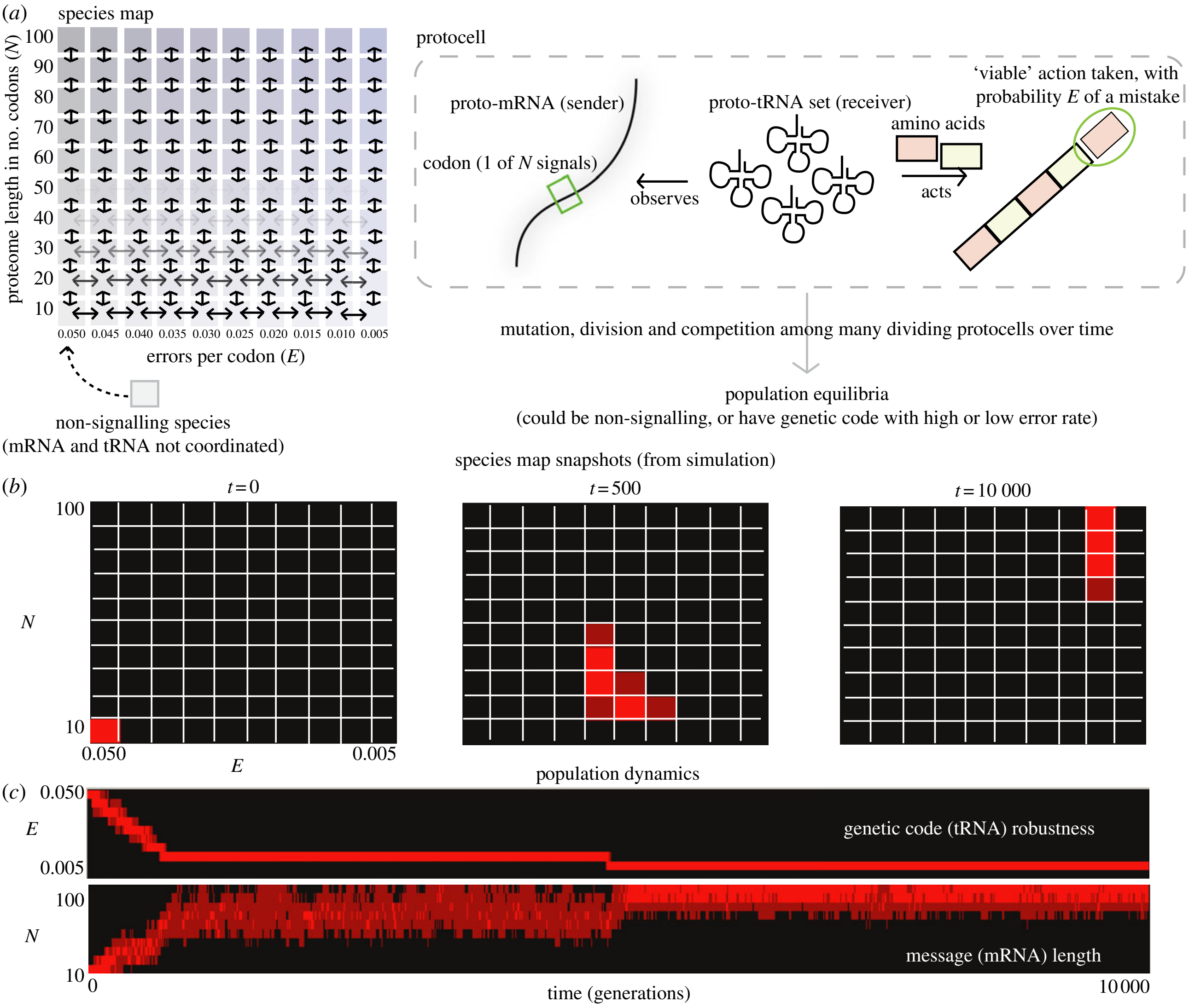

Figure 1. (*a*) Species map. Depicted are the possible species of a given error rate per signal unit, *E*, and messages of length *N*. There are 10 possible lengths (10, 20, . . . ,100 signals) and 10 possible codes with error rates *E* (0.05, 0.045, . . ., 0.005 errors per signal per generation). During reproduction of a protocell of a given species, mutation is permitted between adjacent species on the map. Lighter arrows indicate lower mutation rates (see section Mathematical description of framework). (*b*) 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. (*c*) 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.

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.

We construct a framework for simulating the dynamics of the extended signalling games described above (see section 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 modelled using ordinary differential equations

(ODEs), 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 adenosine triphosphate (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, signalling game theory would suggest that, because of the information asymmetry, after a signalling 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 colocalization 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.

3. 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 signalling convention to dominate the population in a stable manner. Thus, in these simulations, exploration of alternative signalling conventions causes organisms in a population to concentrate towards error-minimal signalling conventions. Similarly, organisms concentrate towards longer messages, but only when the selected signalling convention is robust enough to support them. However, once species of longer message length dominate the population, the adopted signalling 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 both the non-deceptive mRNA and tRNA (figure 2: separate mRNA and tRNA 'syncytia'). Thus, a model based on signalling 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.

4. 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. Although a natural predisposition towards 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 towards 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 signalling 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 signalling 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 codon–anticodon pairs are more thermodynamically stable than other pairs. These pairings likely influenced the development of the genetic code;

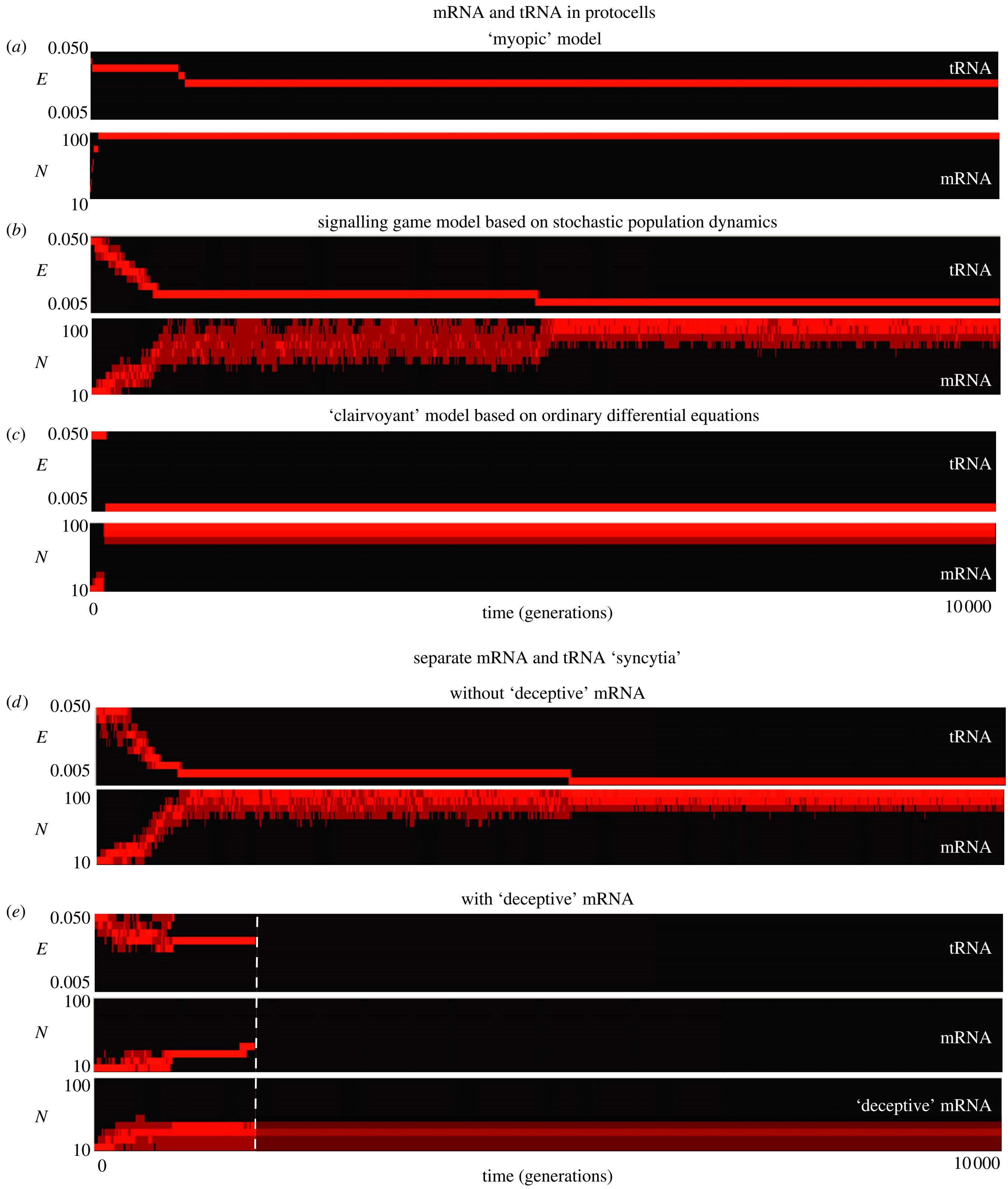

Figure 2. The range of *E*-values (codes represented by tRNA) and *N*-values (message lengths of mRNA) are reported across time from simulations given different assumptions. (*a*) A greedy algorithm for genetic code and genome selection is used as described by other researchers [5,8]. As previously reported, 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 using a greedy approach. We also find that a syncytium 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.

for example, genes that are efficiently expressed take advantage of the bias by over-using 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 protocells 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 signalling model suggests that there existed a reproductive benefit to encoding even short strings of polypeptides, whereas the usage of each codon was still nascent. Even short proteins might have provided structural support to protocells 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 protocell 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 [25–27].

Overall, this paper presents a framework to study signalling 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.

5. Mathematical description of framework

5.1. 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 signalling convention (genetic code) and message (proteome) length. The change in organism number of a species S_i is given by

$$\begin{aligned}\frac{\mathrm{d}S_i}{\mathrm{d}t} = S_i\left(b\left(1-\sum_{j\in\mathcal{L}}\mu_{\text{length}:i\to j}-\sum_{j\in\mathcal{G}}\mu_{c:i\to j}\right)-d\right)\\ +\sum_{j\in\mathcal{G}}S_j\mu_{c:j\to i}+\sum_{j\in\mathcal{L}}S_j\mu_{\text{length}:j\to i},\end{aligned} \tag{5.1}$$

where $\mathcal{L}$ is the set of species with the same genetic code as species i with lengths attainable by one mutation (neighbour lengths); $\mathcal{G}$ is the set of species with the same length as species i with genetic codes attainable by one genetic code mutation (neighbour genetic codes). Although the notation here implies a set of ODEs, 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 rate. In the syncytial model considered in figure 2*d*,*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.

5.2. 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

$$p_{\text{correct}} = \prod_{i\in A}\prod_{j\in A}(1-p_{j\to i})^{k_{j\to i}n_{\text{proteome}:j}}, \tag{5.2}$$

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 $n_{\text{proteome}:\,j}$ 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 of the order of 10^{-3} per codon per translation in yeast [33], and mutation takes place of 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 the sake of 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 $n_{\text{genome}\,:\,\text{tot}}$ is given by

$$b = b_0 p_{\text{correct}} n_{\text{genome}\,:\,\text{tot}}, \tag{5.3}$$

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 $n_{\text{genome}\,:\,\text{tot}}=0$ would correspond to a situation in which proteins encoded by the genome are dysfunctional.

5.3. 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, then 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 the 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

$$\mu_{c\,:\,y\rightarrow x} = \mu_0 \left(\frac{n_{GC:x} - 1}{n_{GC:x}}\right) k_{y\rightarrow x} n_{\text{genome}\,:\,x}, \tag{5.4}$$

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 $n_{\text{genome}\,:\,x}$ is the number of times amino acid x is coded for in the genome. Note that 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 $k_{y\rightarrow x}$ as described in equation (5.1) 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, $n_{\text{genome}\,:\,x}$ decreases accordingly.[4]

5.4. 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

$$d = \frac{\text{total population}}{K}, \tag{5.5}$$

where K is a constant carrying capacity. Negative selection is not explicitly modelled because for small mutation rates, with suitable re-parametrization, 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 $K = 1000$, $k_{y\rightarrow x} = 0.3$, $n_{GC:x} = 2$, $b_0 = 1$, $\mu_{\text{length}} = 0.1$ and $\mu_0 = 1$. In figure 2*e*, 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 electronic supplementary material, figures S1–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 electronic supplementary material).

6. 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/.

Acknowledgements. This work was inspired by discussions with Heeralal Janwa, Michael Wigler, Andi Witzel, Loes Olde Loohuis, Rohit Parikh, and Alfredo Ferro. In particular, B.M. 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. J.J., A.S., S.E.M. and B.M. designed the research. J.J. performed the research. J.J. and B.M. wrote the paper.

Funding statement. This research was funded by two NSF grants nos. NSF CCF-0836649 and NSF CCF-0926166. J.J. was supported by a National Defense Science and Engineering Graduate Fellowship from the US Department of Defense.

Endnotes

[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].

[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 signalling conventions rather than stay in a single separating equilibrium [12,13].

[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.

[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].

References

1. Woese CR. 1965 Order in the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 71–75. (doi:10.1073/pnas.54.1.71)
2. Alff-Sternberger C. 1969 The genetic code and error transmission. *Proc. Natl Acad. Sci. USA* **64**, 584–591. (doi:10.1073/pnas.64.2.584)
3. Crick FHC. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379. (doi:10.1016/0022-2836(68)90392-6)
4. Wong JT. 1975 A co-evolution theory of the genetic code. *Proc. Natl Acad. Sci. USA* **72**, 1909–1912. (doi:10.1073/pnas.72.5.1909)
5. Vetsigian K, Woese C, Goldenfeld N. 2006 Collective evolution and the genetic code. *Proc. Natl Acad. Sci. USA* **103**, 10 696–10 701. (doi:10.1073/pnas.0603780103)
6. Osawa S, Jukes TH. 1989 Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* **28**, 271–278. (doi:10.1007/BF02103422)
7. Massey S. 2008 A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**, 510–516. (doi:10.1007/s00239-008-9167-4)
8. Sella G, Ardell DH. 2006 The coevolution of genes and genetic codes: Crick’s frozen accident revisited. *J. Mol. Evol.* **63**, 297–313. (doi:10.1007/s00239-004-0176-7)
9. Maynard Smith J, Parker GA. 1976 The logic of asymmetric contests. *Anim. Behav.* **24**, 159–175. (doi:10.1016/S0003-3472(76)80110-8)
10. Cho I-K, Kreps DM. 1987 Signaling games and stable equlibria. *Q. J. Econ.* **102**, 179–221. (doi:10.2307/1885060)
11. Skyrms B. 2010 *Signals: evolution, learning and information.* Oxford, UK: Oxford University Press. (doi:10.1093/acprof:oso/9780199580828.001.0001)

12. Jansen VAA, van Baalen M. 2006 Altruism through beard chromodynamics. *Nature* **440**, 663–666. (doi:10.1038/nature04387)
13. Traulsen A, Nowak MA. 2007 Chromodynamics of cooperation in finite populations. *PLoS ONE* **2**, e270. (doi:10.1371/journal.pone.0000270)
14. Smith JM. 1999 The idea of information in biology. *Q. Rev. Biol.* **74**, 395–400. (doi:10.1086/394109)
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. (doi:10.1088/1478-3975/5/1/016001)
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 SE, Garey JR. 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**, 399–410. (doi:10.1007/s00239-005-0260-7)
18. Turk RM, Chumachenko NV, Yarus M. 2010 Multiple translational products from a five-nucleotide ribozyme. *Proc. Natl Acad. Sci. USA* **107**, 4585–4589. (doi:10.1073/pnas.0912895107)
19. Murakami H, Ohta A, Goto Y, Sako Y, Suga H. 2006 Flexizyme as a versatile tRNA acylation catalyst and the application for translation. *Nucleic Acids Symp. Ser.* **50**, 35–36. (doi:10.1093/nass/nrl018)
20. Itzkovits S, Alon U. 2007 The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* **17**, 405–412. (doi:10.1101/gr.5987307)
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**, 199–209. (doi:10.1016/0378-1119(82)90157-3)
22. Weber AL, Pizzarello S. 2006 The peptide-catalyzed stereospecitic synthesis of tetrose: a possible model for prebiotic molecular evolution. *Proc. Natl Acad. Sci. USA* **103**, 12 713–12 717. (doi:10.1073/pnas.0602320103)
23. Mandecki W. 1990 A method for construction of long randomized open reading frames and polypeptides. *Protein Eng.* **3**, 221–226. (doi:10.1093/protein/3.3.221)
24. Omo Y, Kino Y, Sasagawa N, Ishiura S. 2005 Comparative analysis of the cytotoxicity of homopolymeric amino acids. *Biochim. Biophys. Acta* **1748**, 174–179. (doi:10.1016/j.bbapap.2004.12.017)
25. Keefe AD, Szostak JW. 2001 Functional proteins from a random sequence library. *Nature* **410**, 715–718. (doi:10.1038/35070613)
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.* **56**, 162–168. (doi: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.* **58**, 196–202. (doi:10.1007/s00239-003-2542-2)
28. Gerhart J, Kirschner M. 2007 The theory of facilitated variation. *Proc. Natl Acad. Sci. USA* **104**(Suppl. 1), 8582–8589. (doi:10.1073/pnas.0701035104)
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**, 1157–1160. (doi:10.1126/science.1217405)
30. Wright S. 1931 Evolution in mendelian populations. *Genetics* **16**, 97–153.
31. Eigen M. 1971 Self organization of matter and the evolution of biological macromolecules. *Die Naturwissenchafen* **58**, 467–523. (doi:10.1007/BF00623322)
32. Grantham R. 1974 Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864. (doi:10.1126/science.185.4154.862)
33. Kramer EB, Farabaugh PJ. 2006 The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13**, 87–96. (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**, 1577–1585. (doi:10.1126/science.7041255)
35. Fisher RA. 1922 On the dominance ratio. *Proc. Roy. Soc. Edinb.* **42**, 321–341.
36. Schultz DW, Yarus M. 1994 Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* **235**, 1377–1380. (doi:10.1006/jmbi.1994.1094)
37. Silva RM, Miranda I, Moura G, Santos MAS. 2004 Yeast as a model organism for studying the evolution of non-standard genetic codes. *Brief. Funct. Genomics Proteomics* **3**, 35–46. (doi:10.1093/bfgp/3.1.35)