doi:10.1111/j.1558-5646.2010.01074.x

IS LIFE IMPOSSIBLE? INFORMATION, SEX, AND THE ORIGIN OF COMPLEX ORGANISMS

Joel R. Peck[1,2] and David Waxman[1]

[1]*School of Life Sciences, The University of Sussex, Brighton, BN1 9QG, United Kingdom*

[2]*E-mail: J.R.Peck@sussex.ac.uk*

Received December 8, 2009
Accepted June 10, 2010

The earliest organisms are thought to have had high mutation rates. It has been asserted that these high mutation rates would have severely limited the information content of early genomes. This has led to a well-known "paradox" because, in contemporary organisms, the mechanisms that suppress mutations are quite complex and a substantial amount of information is required to construct these mechanisms. The paradox arises because it is not clear how efficient error-suppressing mechanisms could have evolved, and thus allowed the evolution of complex organisms, at a time when mutation rates were too high to permit the maintenance of very substantial amounts of information within genomes. Here, we use concepts from the formal theory of information to calculate the amount of genomic information that can be maintained. We identify conditions under which much higher levels of genomic information can be maintained than previously considered possible among origin-of-life researchers. In particular, we find that the highest levels of information are maintained when many genotypes produce identical phenotypes, and when reproduction occasionally involves recombination between multiple parental genomes. There is a good reason to believe that these conditions are relevant for very early organisms, and thus the results presented may provide a solution to a long-standing logical problem associated with the early evolution of life.

KEY WORDS: **Adaptation, epistasis, models/simulations, mutations, population genetics, sex.**

Eigen's Paradox is a well-known logical problem associated with the origin of complex organisms (Eigen 1971; Eigen and Schuster 1979; Maynard Smith and Szathmáry 1995). Experimental data and logical considerations have led origin-of-life researchers to believe that, early in the history of life, mutation rates were much higher than they are in contemporary organisms (Eigen 1971; Eigen and Schuster 1979; Maynard Smith and Szathmáry 1995). According to Eigen and Schuster, this implies that the maximum amount of information that could have been stably encoded in the genomes of early organisms must have been severely limited (Eigen 1971; Eigen and Schuster 1979). In contemporary organisms, the mechanisms of error prevention and correction are quite complex. This leads to a "chicken-and-egg problem." How could life that is complex enough to suppress mutation to low levels have evolved while mutation rates were quite high?

Eigen and Schuster's calculations are based on the idea that if the genome with the best-possible fitness cannot be maintained in a population, then "The information . . . would slowly seep away until it is entirely lost" (Eigen and Schuster 1979). Using this idea, Eigen and Schuster claimed that, for realistic parameter values, a meaningful genetic sequence cannot include much more than approximately $1/\mu$ nucleotides, where μ is the per-nucleotide mutation rate (Eigen and Schuster 1979). Assuming four equally likely nucleotides, it requires 2 bits of information to specify each nucleotide. Thus, Eigen and Schuster's calculations suggest that the maximum level of biological information that can be stably maintained in the genomes of early organisms is typically of the order of $2/\mu$ bits. Here, the phrase "biological information" refers to the information required by an organism to survive and/or reproduce. A high level of biological information makes high levels of biological complexity possible. That is

to say, it facilitates the maintenance of complex adaptations, including error-correcting mechanisms.

It has been pointed out that the maintenance of biological complexity does not depend on the preservation of any particular genetic sequence; instead, biological complexity depends only on the maintenance of phenotypes that confer relatively high fitness (Huynen et al. 1996; Kun et al. 2005; Takeuchi et al. 2005). However, if phenotypes depend on genotypes, then the biological-information content of the genome can still be calculated. For example, let us assume that every possible phenotype can be uniquely placed into one of Ω distinct categories. Furthermore, each phenotype is produced by a different set of genotypes. Assume that one of these phenotypic categories is associated with a fitness value that is higher than that of any other phenotypic category. In this case, if natural selection leads to individuals of the fittest phenotypic class becoming very common in the population, then from basic information–theoretic considerations (Shannon 1948; Cover and Thomas 1991), it is reasonable to say that the amount of biological information in the genome of a typical population member is $\sim\log_2(\Omega)$ bits (here $\log_2$ denotes a logarithm to base 2). Furthermore, if the only individuals present in the population are members of the fittest phenotypic class, then the genome of every population member can be said to contain exactly $\log_2(\Omega)$ bits of biological information.

To further clarify these ideas, it is useful to consider the case of a phenotype that consists of the amino acid sequence of a protein. Assume the protein is a chain of length A amino acids. If 20 different amino acids can be used in the construction of the protein (as in contemporary organisms) then there are 20^A different possible amino acid sequences of length A. Thus in this case there are $\Omega = 20^A$ distinct phenotypes. If natural selection is sufficiently effective that the only individuals found in a population are members of the fittest phenotypic class, and hence produce proteins with the best possible amino-acid sequence, then, as it takes $\log_2(20^A)$ bits of information to specify this best-possible sequence, it is clearly sensible to say that the genome of each individual contains $\log_2(20^A)$ bits of biological information.

Two studies that used the phenotypic approach to biological complexity, described above, have reported criteria for maintaining a high level of biological complexity that are different from those calculated by Eigen and Schuster. However, for plausible parameter values, these differences were found to be relatively modest (Kun et al. 2005; Takeuchi et al. 2005). Here, we show that there are conditions under which the phenotypic approach leads to the possibility of much more biological information being maintained in a population, for a given mutation rate, than is suggested by Eigen and Schuster's calculations (Eigen 1971; Eigen and Schuster 1979). These conditions include the process of recombination among genomes, and the situation where many genotypes produce the same phenotype. Encouragingly, there is a good reason to think that these conditions are relevant for very early organisms (Huynen et al. 1996; Woese 1998; Wilke 2001; Wilke et al. 2001; Lehman 2003; Santos et al. 2003, 2004; Codoner et al. 2006; Szathmáry 2006; Sanjuan et al. 2007; Soll et al. 2007; Sardanyes et al. 2008).

The idea that, very early in the history of life, recombination occurred between genomes (i.e., sexual reproduction occurred) is absent in the formulations of Eigen and Schuster. However, the possibility of recombination occurring within populations of early organisms has now been widely accepted by the community of origin-of-life researchers (Woese 1998; Lehman 2003; Santos et al. 2003; Santos et al. 2004; Szathmáry 2006; Soll et al. 2007). Sex and recombination can be induced via complex evolved mechanisms, as in many contemporary organisms. However, sex can also result from much more primitive mechanisms, which are more likely to be relevant for the earliest organisms. For example, in the earliest stages of the development of life on Earth, the primordial organisms may have consisted of self-replicating linear polymers that were not contained within a cell membrane. This situation is similar to the process within a polymerase chain reaction (PCR) machine. It is well known that recombination tends to occur during PCR; some partially formed offspring polymers become detached from their parent molecules, and then, as a result of homology, attach to another parent molecule, followed by the completion of replication (Meyerhans 1990). As we will see, the incorporation of sex and recombination can have enormous effects on the amount of information that can be stably maintained in a population in the face of high rates of mutation.

The model we will study here assumes truncation selection. Under truncation selection, an "ideal genotype" exists and any genotype that differs from the ideal genotype by less than a given number of genetic changes is as fit as an individual with the ideal genotype, whereas all individuals with other genotypes have zero fitness. Truncation selection is, essentially, a form of synergistic epistasis, as described by Kondrashov, and others (Kimura and Maruyama 1966; Crow and Kimura 1979; Kondrashov 1988; Peck and Waxman 2000). It is well known that, under synergistic epistasis, sex and recombination can confer large benefits, allowing (in some circumstances) for nearly maximal levels of fitness, despite a high genomic rate of deleterious mutations. This is exactly what occurs in the model presented here.

To the best of our knowledge, the fitness advantages of sex, when the fitness landscape is synergistic, have never previously been invoked in an attempt to solve Eigen's Paradox. At first glance, it may seem certain that synergistic epistasis will be effective in this context. However, there is a complication. Truncation selection implies a certain amount of genetic redundancy, as it occurs only when multiple genotypes exist that all lead to a relatively high level of fitness. However, redundancy decreases the information-carrying capacity of the genome. This is obvious

if we consider the extreme case in which all genotypes generate exactly the same phenotype. In this case mutations can never have a deleterious effect on fitness. However, this does not eliminate Eigen's Paradox, because it is extremely unlikely that the one and only phenotype that can be coded by the genome will happen to be a genome that confers a high level of fitness and/or a low mutation rate.

What we show in this work is that, despite the complication just described, there is an intermediate level of genetic redundancy that can confer a high level of fitness in the face of a high mutation rate, while still allowing for much more information to be stably maintained in the genome than would be expected from Eigen and Schuster's calculations (Eigen 1971; Eigen and Schuster 1979). As we will see, this advantage of genetic redundancy is maximized when genetic material from multiple individuals is often combined, when offspring are produced. That is, the advantages of genetic redundancy are maximized by the occurrence of sexual reproduction.

Our work is closely related to work that has been carried out by a variety of other theorists. These include the biologists mentioned above. Other related findings have been produced by computer scientists, including Baum et al. (1995) and Muhlenbein and Schlierkamp-Vosen (1993). A more closely related study by Watkins (2002) provides substantial insights into the limits of genomic–information content when mutations are common.

The Model

Let us now examine a simple model that was inspired by the models previously investigated by Eigen and Schuster (Eigen 1971; Eigen and Schuster 1979). Consider an organism with a genome that consists of a linear polymer consisting of L monomers. For now, we will assume that there are four different monomer types, as in contemporary organisms. To begin, we follow Eigen and Schuster, and assume that the organism is asexual (Eigen 1971; Eigen and Schuster 1979). We also assume that the population size is very large (effectively infinite, so there are no stochastic effects associated with random genetic drift) and the population size does not change over time. Thus, every individual that "dies" is replaced by the birth of a new organism. (Here "death" includes any mechanism that removes organisms from the reproductive pool, such as denaturing or emigration.)

There are a total 4^L different possible monomer sequences (genomes) that we label 1, 2, ... 4^L. Let d_i represent the death rate for individuals with the ith sequence. Thus, in a very small time interval, Δt, the probability that an individual with genome sequence i will die is $d_i \times \Delta t$. We assume that when a new individual is born, each member of the population is equally likely to be the parent. The new individual has the same genome as its parent, apart from any new mutations it carries. We assume that mutations are independent of one another, and that they occur during the birth process with a probability of μ per monomer. Each mutation has an equal probability of converting the parental monomer to one of the other three monomer types.

To study a population that incorporates sex and recombination, we use exactly the model just specified, except we make the additional assumption that, when a population has come to equilibrium, there is statistical independence of the monomers present at different sites within the genomes of newborn individuals. Thus, when sex and recombination are occurring, we assume that among newborns, the probability of a particular monomer being present, at a particular site within the genome, in a particular individual, is independent of the individual's genotype at all other sites. Technically, this condition can only be guaranteed if each monomer incorporated into an offspring is derived from a different, randomly selected, parent. However, experience in population genetics shows that modest amounts of recombination are typically sufficient to ensure a high level of statistical independence between loci, so long as the population size is large, mating is random, and a large number of sites within the genome are subject to selection (Bulmer 1989; Turelli and Barton 1990; Lynch and Walsh 1998). Whether or not near-statistical independence is a reasonable assumption for a given real population depends on the frequency of recombination events, and on other details that are highly uncertain. We limit attention to the two extremes, ranging from asexuality to complete statistical independence, as a way of measuring the full potential impact of recombination, and to facilitate the calculations.

For the asexual mode of reproduction, we assume that, at equilibrium, the monomer frequencies are the same at all sites within the genome. We make the same assumption for the sexual mode of reproduction. These assumptions are in accord with our experience with numerical studies involving small genomes (low values for L) where we initialized trials using randomly selected frequencies of the various possible genotypes (i.e., generally starting with unequal monomer frequencies at different sites within the genomes). For the main runs, which used larger genomes, we initialized the population with all individuals having a mutation-free genotype.

Our experience in working with the above model indicates that the set of genotypes in the population tends to approach a unique equilibrium distribution. The biological-information content of a typical genome, once this equilibrium distribution is achieved, is our central focus. This issue is easiest to address in the context of a simple scheme for assigning the values of the death rates (the d_i) to genotypes. We now specify this scheme.

SCHEME FOR ASSIGNING DEATH RATES

The phenotypes of our hypothetical organism are classified as falling into Ω different categories, as suggested above. The

categorization is on the basis of all of the phenotypic characteristics that are important for natural selection (i.e., for the determination of the death rates—the d_i). In general, the categorization could be on the basis of multiple phenotypic characteristics, and hence the value of Ω can be extremely large (Waxman and Welch 2005).

Let us assume that, in a particular environment, only one of the Ω different possible phenotypes will allow survival. We will call this phenotype the "high-fitness phenotype." (Our investigations have shown that this assumption is not a crucial determinant of the qualitative nature of the results. However, it does simplify matters). We measure time in units where individuals with the high-fitness phenotype have a death rate of unity ($d_i = 1$); all other phenotypes have an infinite death rate ($d_i = \infty$). It follows that only genomes that produce the high-fitness phenotype will be found in the population. This means that the genome of any individual present in the population encodes $\log_2(\Omega)$ bits of biological information.

For simplicity, let us assume that each possible phenotypic category (of which there are a total of Ω) is generated by a unique set of $4^L/\Omega$ different genomes (thus, $4^L/\Omega$ must be an integer). Additionally, we assume that all genomes that produce the high-fitness phenotype have a similar sequence. In particular, we assume that a particular "ideal" genome can be identified, and this genome produces the high-fitness phenotype. We assume further that the high-fitness phenotypes is also produced by all genomes that have a sequence with, at most, θL differences from the sequence of the ideal genome (where $0 \leq \theta \leq 1$). For example, say that $\theta = 0.4$ and $L = 10{,}000$. Under these conditions, our assumptions imply that the high-fitness phenotype in question will be produced by any genome that differs from the ideal genotype at 4000 locations within the genome, or fewer.

Results

Consideration of the model suggests that there is no upper limit to the amount of biological information than can be maintained in a population, as long as the genome is sufficiently large. For example, if each phenotypic category is produced by only a single genomic sequence ($\theta = 0$) then the number of phenotypic categories equals the number of different genotypes, that is, $\Omega = 4^L$ and so $\log_2(4^L) = 2L$ bits of biological information can be maintained in an organism. The amount of biological information that can be maintained is smaller with larger values of θ, but the amount of maintainable information always increases with L for any value of θ. This observation, however, does not lead to a realistic solution to Eigen's Paradox unless we can show that the information can be maintained without imposing a high "genetic load" on the population. In the current context, this means establishing that the average number of offspring produced by adults is not unreasonably large. (Here, an "adult" is an individual that possesses the high-fitness phenotype, and thus survives the birth process.) With these considerations in mind, we focus on the birth rate of adults; this is a quantity that, in a population whose number is regulated at equilibrium, has a value determined by parameters of the model, and by the mode of reproduction.

We will use B to denote the equilibrium birth rate. We measure B in the time units adopted above, where the mean lifetime of individuals with the high-fitness phenotype is unity. Thus, if an equilibrium population contains N adults, then, during a period of time equal to an average adult lifetime (i.e., during one time unit) approximately NB offspring will be born. We numerically determine the equilibrium birth rate from the equilibrium distribution describing a population. This is arrived at from consideration of the long-term dynamics of a population, and the condition for equilibrium leads to an equation that can be numerically iterated to determine the equilibrium distribution (see Appendix 1).

Figure 1 shows the value of the equilibrium birth rate, B, for various values of the parameter θ, when the per-monomer mutation rate is $\mu = 0.01$. The figure also shows the amount of biological information present in the genomes of these adults, namely $\log_2(\Omega)$ bits. A much more comprehensive set of results is given in Table 1.

We note that to calculate Ω (and thus information content) we use θ and L to calculate the proportion of all possible genomes that produce the high-fitness genotype, and thus allow survival. The reciprocal of this proportion is equal to Ω (see Appendix 1).

The results reveal that for the stable encoding of a given number of bits of biological information, the equilibrium birth rate, under sexual reproduction, is lower than that required under asexual reproduction (and often much lower). The only exception to this occurs when there is no genetic redundancy, so every possible genome leads to a different phenotype, and all alterations to the ideal genetic sequence are fatal (i.e., when, $\theta = 0$ so $\Omega = 4^L$). In this case, the equilibrium birth rate required for survival of an asexual population is identical to that required by a sexual population.

Discussion

Is the amount of information that can be stably encoded in a pre-enzymatic world sufficient to produce the sort of complex phenotypes that could reduce error rates, thereby allowing more complex phenotypes, and yet lower error rates, etc? To address this question, it is useful to note that even if every offspring born produces the high-fitness phenotype, the equilibrium birth rate can never fall below $B = 1$, given the assumption of an unchanging population size. Therefore, a birth rate that is no

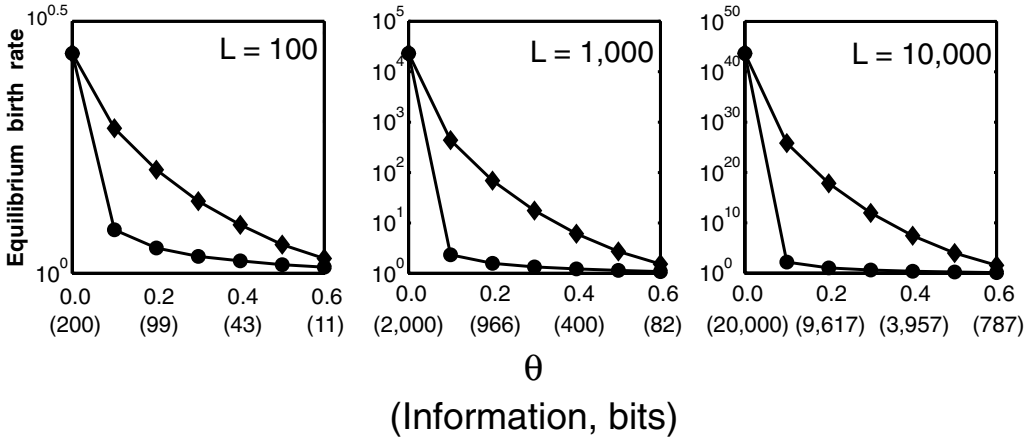

Figure 1. Equilibrium birth rates as a function of θ when μ = 0.01 (where μ is the rate of mutation per monomer). The three panels are for three different values of the number of monomers in the genome, *L*. The ordinates give the equilibrium birth rate (i.e., the mean number of offspring produced per adult) for different parameter values. The diamonds indicate results for asexual reproduction, whereas filled circles indicate results for sexual reproduction. The rows of numbers just beneath the abscissas give values of θ, which denotes the proportion of monomers that must be identical to those in the "ideal" genome if the organism is to survive. The lower row of numbers gives the information values in bits (binary digits).

more than 20 times this absolute minimum value does not seem implausibly high (i.e., $B \leq 20$). Indeed, much higher birth rates are not out of the question. In the absence of error-correcting enzymes, mutation rates might have been as low as $\mu = 0.01$ (Eigen 1971; Maynard Smith and Szathmary 1995). With these considerations in mind, it is of interest to note (from Fig. 1 and Table 1) that with a mutation rate of $\mu = 0.01$ and a birth rate below 12, it is possible to encode 9617 bits of information when reproduction is sexual and the genome length is $L = 10{,}000$. This is much more information than would be expected using Eigen and Schuster's calculations (Eigen 1971; Eigen and Schuster 1979), which lead to the expectation that not much more than $2/\mu = 200$ bits of biological information can be maintained when $\mu = 0.01$ and birth rates are not extremely large. Thus, Eigen's Paradox appears to be much less paradoxical.

To understand this finding in more concrete terms, it may help to recognize that if the phenotype being specified by the genome is a protein sequence constructed from 20 amino acids then 9617 bits is sufficient to provide an exact specification of a protein sequence that is 2225 amino acids in length ($9617 \approx \log_2(20^{2225})$). If the phenotype consists of a ribozyme constructed from four nucleotides, then 9617 bits can specify a sequence of 4808 nucleotides ($9,617 \approx \log_2(4^{4808})$).

Do our findings represent a key step in solving Eigen's Paradox? This depends (in part) on the minimum required complexity of an effective error-reducing agent (e.g., an error-reducing enzyme or ribozyme). Relevant data are not yet at a stage when a definitive statement on this matter can be made. However, data that bear directly on this issue, come from the laboratory production of a ribozyme that replicates short RNA sequences with an error rate of $\mu \approx 0.033$. This ribozyme is only 189 nucleotides in length, and thus requires $\log_2(4^{189}) = 378$ bits of information to specify the complete sequence. The ribozyme was developed in the course of a relatively short experiment (Johnston et al. 2001). Thus, it would not be surprising if a much less-mutagenic ribozyme could be specified with no more data, and perhaps even less. However, even if the ribozyme would have to be twice as long to be much less mutagenic, the required amount of information ($2 \times 378 = 756$ bits) can be stably maintained for a mutation rate that is as high as $\mu = 0.04$ (see Table 1 for a sexual population with $L = 10{,}000$ and $\theta = 0.6$). Furthermore, this can be accomplished without requiring a high birth rate. These observations, although not definitive, are at least suggestive that the combined chemical and mathematical realities do not conspire to make life impossible.

As noted above, our model incorporates truncation selection, which is, essentially, a form of synergistic epistasis (Kimura and Maruyama 1966; Kondrashov 1988; Peck and Waxman 2000). Synergistic epistasis means that deleterious mutations tend to be more damaging in genomes that are already contaminated with many deleterious mutations, as compared to their effects in less-contaminated genomes. Although the use of truncation selection helps to simplify and clarify our presentation, it is certainly not necessary to generate results that are qualitatively similar to ours. Nevertheless, it is important to recognize that, for selection schemes that do not incorporate synergistic epistasis, results that are very different from ours are likely to emerge. In particular, it is straightforward to find cases in which, in contrast to our results, recombination between genomes increases the birth rate required to sustain the population at equilibrium.

Table 1. The per-monomer mutation rates (μ) and mode of reproduction (asexual or sexual) are specified in the left-most column, and the data to the right of these specifications give the mean number of offspring for adult members of a population at equilibrium under the given parameter values. The question marks denote data points too large to calculate using our methods. The quantity *L* gives the number of monomers in the genome, and θ denotes the proportion of monomers that must be identical to those in the "ideal" genome if the organism is to survive.

θ	0	0.1	0.2	0.3	0.4	0.5	0.6
L=100							
Info. (bits)	200.00	140.00	99.00	68.00	43.00	24.00	11.00
μ=0.01, asexual	2.73	1.94	1.61	1.39	1.25	1.14	1.07
μ=0.01, sexual	2.73	1.22	1.12	1.08	1.06	1.04	1.03
μ=0.02, asexual	7.54	3.77	2.59	1.95	1.56	1.31	1.15
μ=0.02, sexual	7.54	1.58	1.31	1.20	1.14	1.09	1.06
μ=0.04, asexual	59.28	14.58	6.78	3.82	2.44	1.71	1.31
μ=0.04, sexual	59.28	3.06	1.91	1.54	1.35	1.23	1.14
μ=0.08, asexual	4.18×10^{3}	235.43	49.11	15.21	6.06	2.95	1.72
μ=0.08, sexual	4.18×10^{3}	17.6	5.31	2.98	2.08	1.62	1.34
L=1,000							
Info. (bits)	2,000.00	1,377.00	966.00	648.00	400.00	212.00	82.00
μ=0.01, asexual	2.32×10^{4}	439.23	68.88	17.58	6.11	2.72	1.52
μ=0.01, sexual	2.32×10^{4}	2.33	1.58	1.34	1.22	1.14	1.08
μ=0.02, asexual	5.94×10^{8}	2.05×10^{5}	4.94×10^{3}	316.87	37.81	7.43	2.32
μ=0.02, sexual	5.94×10^{8}	10.29	3.36	2.13	1.63	1.37	1.19
μ=0.04, asexual	5.36×10^{17}	5.56×10^{10}	2.94×10^{7}	1.13×10^{5}	1.53×10^{3}	57.00	5.44
μ=0.04, sexual	5.36×10^{17}	999.27	33.40	8.25	3.75	2.23	1.53
μ=0.08, asexual	1.63×10^{36}	1.05×10^{22}	1.98×10^{15}	2.19×10^{10}	3.22×10^{6}	3.80×10^{3}	31.03
μ=0.08, sexual	1.63×10^{36}	1.90×10^{9}	6.33×10^{4}	671.13	50.66	9.58	3.05
L=10,000							
Info. (bits)	20,000.00	13,731.00	9,617.00	6,439.00	3,957.00	2,082.00	787.00
μ=0.01, asexual	4.45×10^{43}	6.70×10^{25}	6.98×10^{17}	9.37×10^{11}	2.77×10^{7}	9.98×10^{3}	37.26
μ=0.01, sexual	4.45×10^{43}	164.27	11.66	4.30	2.49	1.73	1.33
μ=0.02, asexual	5.49×10^{87}	9.40×10^{51}	8.06×10^{35}	1.22×10^{24}	9.36×10^{14}	1.10×10^{8}	1.43×10^{3}
μ=0.02, sexual	5.49×10^{87}	2.19×10^{7}	2.50×10^{3}	85.19	14.12	4.56	2.10
μ=0.04, asexual	1.94×10^{177}	1.95×10^{105}	5.39×10^{72}	6.02×10^{48}	2.04×10^{30}	1.86×10^{16}	2.39×10^{6}
μ=0.04, sexual	1.94×10^{177}	2.02×10^{25}	5.91×10^{11}	3.27×10^{6}	5.34×10^{3}	105.85	8.24
μ=0.08, asexual	?	2.05×10^{216}	2.50×10^{149}	1.47×10^{100}	1.62×10^{62}	2.32×10^{33}	1.13×10^{13}
μ=0.08, sexual	?	8.05×10^{85}	2.23×10^{42}	4.03×10^{23}	1.94×10^{13}	8.73×10^{6}	812.63

Is it likely that synergistic epistasis was a common mode of selection among the early ancestors of life on Earth? Although a comprehensive discussion of this question is beyond the scope of the current work, it is worth noting that both computer simulations and experimental data suggest that synergistic epistasis tends to arise when genomic mutation rates are relatively high, as they presumably were during the early stages of the development of complex life (Wilke 2001; Wilke et al. 2001; Proulx and Phillips 2005; Codoner et al. 2006; Sanjuan et al. 2007; Sardanyes et al. 2008). This makes sense, as redundancy in the production of phenotypes is likely to be important when genomic mutation rates are high, as redundancy means that damage to one region of the genome can be ameliorated by other, undamaged regions. This sort of redundancy tends to generate synergistic epistasis (Kondrashov 1988; Sanjuan and Elena 2006). In addition, theoretical work suggests that synergistic epistasis tends to arise easily when selection depends on competitions between small numbers of individuals (Hamilton and Tanese 1990; Peck and Waxman 2000). Early organisms might have competed, for example, for attachment sites in which they could anchor to a substrate, or for monomers from which to produce the next generation. Finally, it may be that the development of complex life simply had to wait until the appearance of a selection regime that was sufficiently synergistic. Such an appearance does not seem unlikely, given that an appropriate selection regime need only have been present within a limited area, and only for a period of time sufficient for the development of efficient mechanisms for the reduction of the mutation rate.

In the calculations presented here, we have assumed that the number of possible monomers is four—as in contemporary organisms. However, this need not have been the case early in the history of life. The maximum amount of information that can be encoded by a genetic sequence of a given length increases with the number of monomers. This maximum is achieved when $\theta = 0$, so any difference from the "ideal" genetic sequence is fatal, and every genome codes for a different phenotype. In this case $\Omega = m^L$, where m is the number of monomers in use. Thus, the maximum amount of information achievable (expressed in bits) is given by $\log_2(m^L) = L \log_2(m)$. This expression shows that, although more information can be encoded by using more monomers, the increase is only logarithmic in the number of monomers, m, and hence quite insensitive to this number.

Our results are, of course, dependent on the details of the model that we chose. Our calculations should, therefore, be regarded as an "existence proof." That is, they show that when recombination occurs, simple coding and selection schemes exist that allow for much more information to be encoded into genomes than was believed possible by the framers of Eigen's Paradox (Eigen 1971; Eigen and Schuster 1979; Maynard Smith and Szathmary 1995). We hope that this observation will lead to new and productive avenues of research among origin-of-life researchers, and perhaps among researchers concerned with contemporary organisms as well.

ACKNOWLEDGMENTS

We thank G. F. Joyce, J. Maynard Smith, L. E. Orgel, E. Szathmáry, C. J. C. H. Watkins, J. Welch, D. Woolfson, and two anonymous reviewers for helpful discussions and comments. This work was supported by the Biotechnology and Biological Sciences Research Council (UK), The Leverhulme Trust, and the Centre for Computational Systems Biology, Fudan University, Shanghai.

LITERATURE CITED

Baum, E. B., D. Boneh, and C. Garrett. 1995. On genetic algorithms. *Proceedings of the Eighth Annual Conference on Computational Learning Theory* 230–239.

Bulmer, M. G. 1989. Maintenance of genetic variability by mutation-selection balance: a child's guide through the jungle. Genome 31:761–767.

Codoner, F. M., J. A. Daros, R. V. Sole, and S. F. Elena. 2006. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. PLOS Pathogens 2:1187–1193.

Cover, T. M., and J. A. Thomas. 1991. Elements of information theory. Wiley-Interscience, New York.

Crow, J. F., and M. Kimura. 1979. Efficiency of truncation selection. Proc. Natl. Acad. Sci. USA 76:396–399.

Eigen, M. 1971. Self organization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523.

Eigen, M., and P. Schuster. 1979. The hypercycle: a principle of natural self-organization. Springer, Berlin.

Hamilton, W. D., R. Axelrod, and R. Tanese. 1990. Sexual reproduction as an adaptation to resist parasites (a review). Proc. Natl. Acad. Sci. USA 87:3566–3573.

Huynen, M. A., P. F. Stadler, and W. Fontanta. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. Proc. Natl. Acad. Sci. USA 93:397–401.

Johnston, W. K., P. J. Unrau, M. S. Lawrence, M. F. Glasner, and D. P. Bartel. 2001. RNA-catalyzed RNA- polymerization: accurate and general RNA-templated primer extension. Science 292:1319–1325.

Kimura, M., and T. Maruyama. 1966. The mutational load with epistatic gene interactions. Genetics 54:1337–1351.

Kondrashov, A. S. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435–440.

Kun, A., M. Santos, and E. Szathmáry. 2005. Real ribozymes suggest a relaxed error threshold. Nat. Genet. 37:1008–1111.

Lehman, N. 2003. The case for the extreme antiquity of recombination. J. Mol. Evol. 56:770–777.

Lynch, M., and J. B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Assoc., Inc., Sunderland, MA.

Maynard Smith, J., and E. Szathmáry. 1995. The major transitions in evolution. W. H. Freeman Spektrum, Oxford.

Meyerhans, A., J. P. Vartanian, and S. Wain-Hobson. 1990. DNA recombination during PCR. Nucleic Acids Res. 18:1687–1691.

Muhlenbein, H., and D. Schlierkamp-Vosen. 1993. Predictive models for the breeder genetic algorithm: 1. Continuous parameter optimisation. Evol. Comput. 1:25–50.

Peck, J. R., and D. Waxman. 2000. Mutation and sex in a competitive world. Nature 406:399–404.

Proulx, S. R., and P. C. Phillips. 2005. The opportunity for canalization and the evolution of genetic networks. Am. Nat. 165:147–162.

Sanjuan, R., J. M. Cuevas, V. Furio, and E. C. Holmes. 2007. Selection for robustness in mutagenized RNA viruses. PLOS Genet. 3:939–946.

Sanjuan, R., and S. F. Elena. 2006. Epistasis correlates to genomic complexity. Proc. Natl. Acad. Sci. USA 103:14402–154405.

Santos, M., E. Zintzaras, and E. Szathmáry. 2003. Origin of sex revisited. Origins of Life and Evolution of the Biosphere 33:405–432.

———. 2004. Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizable genome. J. Mol. Evol. 59:507–519.

Sardanyes, J., S. F. Elena, and R. V. Sole. 2008. Simple quasispecies models for the survival-of-the-flattest effect: the role of space. J. Theor. Biol. 250:560–568.

Shannon, C. E. 1948. A mathematical theory of communication. Bell Syst. Techn. J. 27:379–423 and 623–656.

Soll, S. J., C. D. Arenas, and N. Lehman. 2007. Accumulation of deleterious mutations in small abiotic populations of RNA. Genetics 175:267–275.

Strang, G. 1988. Linear algebra and its applications. Harcourt Brace, San Diego.

Szathmáry, E. 2006. The origin of replicators and reproducers. Philos. Trans. R. Soc. Lond. B. 361:1761–1776.

Takeuchi, N., P. H. Poorthuis, and P. Hogeweg. 2005. Phenotypic error threshold; additivity and epistasis in RNA evolution. BMC Evol. Biol. 5:9.

Turelli, M., and N. H. Barton. 1990. Dynamics of polygenic characters under selection. Theor. Pop. Biol. 38:1–57.

Watkins, C. J. C. H. 2002. The channel capacity of evolution: ultimate limits on the amount of information maintainable in the genome. *Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure* 2:58–60.

Waxman, D., and J. J. Welch. 2005. Fisher's microscope and Haldane's ellipse. Am. Nat. 166:447–457.

Wilke, C. O. 2001. Selection for fitness vs. selection for robustness in RNA secondary structure folding. Evolution 55:2412–2420.

Wilke, C. O., J. R. Wang, C. Ofria, R. E. Lenski, and C. Adami. 2001. Evolution of digital organisms at high mutation rate leads to survival of the flattest. Nature 412:331–333.

Woese, C. R. 1998. The universal ancestor. Proceedings of the National Academy of Sciences (USA) 95:6854–6859.

Associate Editor: J. Hermisson

Appendix 1

In this Appendix, we give (1) the formulation and mathematical analysis leading to the results we use for an asexually reproducing population of polymers, (2) the corresponding formulation and analysis for a sexual population, and (3) the way information is calculated for this work.

(1) Asexual reproduction

Consider polymers consisting of L monomers that are located at L contiguous sites. We assume there are m different types of monomer. We will refer to the polymers as "individuals." Each individual is characterized by a vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ where each α_i can take only two values: 0 and 1. When $\alpha_i = 0$ the optimal monomer is present at site i, whereas, when $\alpha_i = 1$, one of the other $m - 1$ suboptimal monomers is present at site i. We will refer to $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ as the "genotype" of an individual. The number of suboptimal monomers (also termed mutations) associated with genotype α, which we write as $n(\alpha)$, is given by

$$n(\alpha) = \sum_{i=1}^{L} \alpha_i. \tag{A1}$$

The dynamics in continuous time of an effectively infinite population of polymers is given in terms of a probability distribution, $\varphi(\alpha, t)$. This is the proportion of the population that has genotype α at time t. To determine the dynamical behavior of a population of polymers, we consider the events that occur in an infinitesimal time interval, namely death of some individuals and the asexual production of offspring by other individuals. This leads to the equation

$$\begin{aligned}\frac{d\varphi(\alpha, t)}{dt} &= -D(n(\alpha))\varphi(\alpha, t) + B(t)\,\chi_A(\alpha, t) \\ &\quad - \varphi(\alpha, t)[-\bar{D}(t) + B(t)],\end{aligned} \tag{A2}$$

where (1) $D(n(\alpha))$ is the death rate of individuals with genotype α; it is assumed to be solely a function of the number of suboptimal monomers (mutations), $n(\alpha)$, associated with genotype α. (2) $B(t)$ is the birth rate (rate of production) of "offspring" polymers at time t. The birth rate of an individual is assumed to be independent of their genotype. (3) $\bar{D}(t)$ is the mean death rate of the population at time t and is given by $\bar{D}(t) = \sum_\alpha D(n(\alpha))\,\varphi(\alpha, t)$ where here and elsewhere we use the notation $\sum_\alpha = \sum_{\alpha_1=0}^{1} \sum_{\alpha_2=0}^{1} \cdots \sum_{\alpha_L=0}^{1}$. (4) $\chi_A(\alpha, t)$ is the probability distribution of newly born individuals. It has the form $\chi_A(\alpha, t) = \sum_\beta M_{\alpha_1\beta_1} M_{\alpha_2\beta_2} \ldots M_{\alpha_L\beta_L} \varphi(\beta, t)$ where $M_{\alpha\beta}$ is the probability that a parental site with monomer of type β (=0 or 1) will, on reproduction, produce an offspring with monomer of type α (=0 or 1) at that site. We have

$$M_{00} = 1 - \mu, \quad M_{01} = \nu, \quad M_{10} = \mu, \quad M_{11} = 1 - \nu, \tag{A3}$$

where $\mu(\nu)$ is the probability that an optimal (suboptimal) monomer will be reproduced as a suboptimal (optimal) monomer. We take $\nu = \mu/(m - 1)$, assuming that mutation rates are identical between different monomers, so given a suboptimal monomer undergoes mutation, it has a probability of $1/(m - 1)$ of producing an optimal monomer.

We work under the assumption that population number is regulated, so the number of polymers present in the population is very large but has a fixed value. As a consequence, the birth rate must equal the mean death rate at all times: $B(t) = \bar{D}(t)$, in which case equation (A2) reduces to

$$\frac{d\varphi(\alpha, t)}{dt} = -D(n(\alpha))\,\varphi(\alpha, t) + \bar{D}(t)\chi_A(\alpha, t). \tag{A4}$$

SYMMETRIC SOLUTIONS

We restrict all considerations to a class of distributions we refer to as "symmetric." Such distributions have the property that all genotypes with k mutations are present in equal proportions in the population. For example, when there are $L = 3$ sites, the genotypes with $k = 1$ mutations, namely (1, 0, 0), (0, 1, 0), and (0, 0, 1) are all present in equal proportions in the population. Numerical solution of equation (A4) suggests that a symmetric distribution always arises at long times. There is a computational advantage to considering symmetric distributions. Instead of having to consider 2^L different genotypes, as we would in a general distribution, a symmetric distribution, involves only $L + 1$ different classes of the number of mutations.

To define a symmetric distribution, let $\psi(k, t)$ be the proportion of the population with k mutations at time t. With $\delta_{a,b}$ a Kronecker delta ($\delta_{a,b}$ has the value 1 when $a = b$ and vanishes otherwise) we can write $\psi(k, t) = \sum_\alpha \delta_{n(\alpha),k}\varphi(\alpha, t)$ for $k = 0, 1, 2, \ldots, L$. With $\binom{L}{k} = \frac{L!}{(L-k)!k!}$ a binomial coefficient, there are $\sum_\alpha \delta_{n(\alpha),k} = \binom{L}{k}$ different genotypes with k mutations. In general, all $\binom{L}{k}$ of these different genotypes will be present in the population in different proportions. A symmetric distribution corresponds to all of these genotypes each being present in the proportion $\psi(k, t)/\binom{L}{k}$. In general, a symmetric distribution is defined by

$$\varphi(\alpha, t) = \frac{\psi(n(\alpha), t)}{\binom{L}{n(\alpha)}}. \tag{A5}$$

For such a distribution, all relevant information about the population is contained in the distribution of mutation numbers, $\psi(k, t)$. Emphasis can thus be shifted from $\varphi(\alpha, t)$ to $\psi(k, t)$. The dynamical equation obeyed by $\psi(k, t)$ follows from equation (A4) by multiplying this equation by $\delta_{n(\alpha),k}$ and summing over all α. This leads to

$$\frac{d\psi(k,t)}{dt} = -D(k)\psi(k,t) + \bar{D}(t)\sum_{j=0}^{L} Q(k,j)\psi(j,t), \quad \text{(A6)}$$

where $\bar{D}(t) = \sum_{k=0}^{L} D(k)\psi(k,t)$ and

$$Q(k,j) = \sum_{x=\max(0,k+j-L)}^{\min(k,j)} \binom{L-j}{k-x}\binom{j}{x}(1-\nu)^x \nu^{j-x}\mu^{k-x}(1-\mu)^{L-k-j+x}. \quad \text{(A7)}$$

EFFECTS OF TRUNCATION SELECTION AT EQUILIBRIUM

In this work, we will consider only equilibrium properties of a population. The proportion of the population with k mutations at equilibrium is denoted $\psi(k)$. It follows from solving a time-independent analogue of equation (A6), namely

$$-D(k)\psi(k) + \bar{D}\sum_{j=0}^{L} Q(k,j)\psi(j) = 0, \quad \text{(A8)}$$

where $\bar{D}$ is the equilibrium mean death rate. We consider death rates given by

$$D(k) = \begin{cases} D_0, & k \le n^* \\ D_1, & k > n^*, \end{cases} \quad \text{(A9)}$$

where n^* is a positive integer and the parameter D_1 is larger than D_0. In the limit $D_1 \longrightarrow \infty$, we have truncation selection where having more than n^* mutations is lethal.

There is a subtlety about calculating the equilibrium mean death rate, $\bar{D}$, in the limit $D_1 \to \infty$, because the fraction of the population with more than n^* mutations is numerically found to be proportional to $1/D_1$. As a consequence, although this fraction of the population, namely $\sum_{k>n^*}\psi(k)$, becomes vanishingly small when $D_1 \to \infty$, it generally will make a finite contribution to $\bar{D}$. This arises because $\bar{D}$ contains the contribution $D_1\sum_{k>n^*}\psi(k)$. A robust way to calculate $\bar{D}$ is to note equation (A8) can be written as

$$\psi(k) = \bar{D}\,[D(k)]^{-1}\sum_{j=0}^{L} Q(k,j)\psi(j). \quad \text{(A10)}$$

Summing this equation over all k and using $\sum_{k=0}^{L}\psi(k) = 1$ yields

$$\bar{D} = \left[\sum_{k=0}^{L}[D(k)]^{-1}\sum_{j=0}^{L} Q(k,j)\psi(j)\right]^{-1}. \quad \text{(A11)}$$

In this form, we can harmlessly take $D_1 \to \infty$ because $\sum_{j=0}^{L} Q(k,j)\psi(j)$ is bounded (<1). The equilibrium mean death rate, $\bar{D}$, equals the equilibrium birth rate of the population, B, because population number was assumed to be unchanging.

We note that at all sites we have included mutations both from and to the optimal monomer. If we neglect mutations from suboptimal monomers to optimal monomers (i.e., neglect "back mutations"), by setting $\nu = 0$ in equation (A3) then we have $Q(k,j) = \binom{L-j}{k-j}(1-\mu)^{L-k}\mu^{k-j}$. In this case, setting $k = 0$ in equation (A10) and assuming $\psi(0) \neq 0$ yields the standard result $\bar{D} = (1-\mu)^{-L}D(0) \simeq e^{\mu L}D(0)$. Including "back mutations" leads to a different expression for $\bar{D}$.

ITERATIVE APPROACH TO AN EQUILIBRIUM SOLUTION

Using equation (A11) to eliminate $\bar{D}$ from equation (A10) yields

$$\psi(k) = \frac{[D(k)]^{-1}\sum_{j=0}^{L} Q(k,j)\psi(j)}{\sum_{g,j=0}^{L}[D(g)]^{-1}Q(g,j)\psi(j)}. \quad \text{(A12)}$$

This suggests that iterating this equation is a possible procedure to determine $\psi(k)$. That is, given a form for $\psi(k)$, we calculate the right-hand side of the equation, to produce an updated form for $\psi(k)$. We repeat this procedure to convergence. This has been found to work in practice for truncation selection when the initial form for $\psi(k)$ corresponds to a mutation-free population: $\psi(k) = \delta_{k,0}$. Once the converged form for $\psi(k)$ has been determined, the equilibrium mean death rate, $\bar{D}$, may be found using the converged $\psi(k)$ in equation (A11).

When $0 < \mu < 1$, all of the elements of the matrix $Q(k, j)$ are positive. It follows that with "near truncation selection" (D_1 large but $<\infty$), the matrix appearing in the iteration scheme described above, namely $R_{k,j} = [D(k)]^{-1}Q(k,j)$, has positive elements, and the iteration is guaranteed to converge, by the Perron Frobenius theorem (Strang 1988).

(2) Sexual population

We will consider the equilibrium properties of a sexual population, under the assumption that offspring consist of L monomers derived from L randomly picked individuals. The analogue of equation (A4), for a sexual population, is

$$\frac{d\varphi(\alpha,t)}{dt} = -D\,(n(\alpha))\,\varphi(\alpha,t) + \bar{D}(t)\chi_S(\alpha,t). \quad \text{(A13)}$$

Here $\chi_S(\alpha, t)$ is the probability distribution of newly born individuals. It has the form

$$\chi_S(\alpha, t) = \sum_{\beta_1=0}^{1} M_{\alpha_1\beta_1}\varphi_1(\beta_1, t) \times \sum_{\beta_2=0}^{1} M_{\alpha_2\beta_2}\varphi_2(\beta_2, t)\ldots\sum_{\beta_L=0}^{1} M_{\alpha_L\beta_L}\varphi_L(\beta_L, t), \tag{A14}$$

where the $M_{\alpha\beta}$ are given in equation (A3) and $\varphi_i(\beta, t)$ is the marginal distribution of monomers of site i: $\varphi_i(\beta_i, t) = \sum_\gamma \delta_{\beta_i,\gamma_i}\varphi(\gamma, t)$.

Numerical solution of equation (A13) suggests that, at long times, $\varphi(\alpha, t)$ settles down to a solution where the marginal distributions associated with each site (the $\varphi_i(\beta_i, t)$) become identical. That is, at large t,

$$\varphi_1(\beta, t) = \varphi_2(\beta, t)\ldots = \varphi_L(\beta, t). \tag{A15}$$

Because we are interested in long time (equilibrium) properties of a population, we will incorporate equality of the marginal distributions directly into the problem.

The distribution of mutations in the population at time t, written $\psi(k, t)$, is defined by $\psi(k, t) = \sum_\alpha \delta_{n(\alpha),k}\varphi(\alpha, t)$. The dynamical equation for $\psi(k, t)$ follows by multiplying equation (A13) by $\delta_{n(\alpha),k}$ and summing over all α. Using equation (A15) we obtain

$$\frac{d\psi(k,t)}{dt} = -D(k)\psi(k,t) + \bar{D}(t)\binom{L}{k}[1 - A(t)]^k\,[A(t)]^{L-k}, \tag{A16}$$

where $A(t) = (1-\mu)[1-p(t)] + \nu p(t)$ and $p(t) = L^{-1}\sum_{k=0}^{L} k\psi(k, t)$.

We adopt a similar procedure to that used in the asexual case to determine the equilibrium solution, by rewriting the equilibrium form of equation (A16) as

$$\psi(k) = \frac{[D(k)]^{-1}\binom{L}{k}(1-A)^k\,A^{L-k}}{\sum_{j=0}^{L}[D(j)]^{-1}\binom{L}{j}(1-A)^j\,A^{L-j}} \tag{A17}$$

and iterating this equation to convergence. At convergence, equation (A16) implies that the equilibrium mean death rate depends on $\psi(k)$ only via its dependence on A, and is given by $\bar{D} = [\sum_{k=0}^{L}[D(k)]^{-1}\binom{L}{k}(1-A)^k A^{L-k}]^{-1}$.

(3) Information content

In the final part of this Appendix, we give details leading to the determination of the information content when: (a) polymers have L sites, with m monomers at each site, (b) the "ideal sequence" has specific monomers at each site, (c) all sequences found in a population have $\leq n^*$ differences from the ideal sequence.

To determine the information content, we note that the total number of possible sequences is m^L and that the number of sequences with k differences from the ideal sequence is $(m-1)^k\binom{L}{k}$. It then follows that the proportion of all sequences with $\leq n^*$ differences from the ideal sequence (defined in the main text as Ω^{-1}), is given by

$$\Omega^{-1} = \frac{\sum_{k=0}^{n^*}(m-1)^k\binom{L}{k}}{m^L}. \tag{A18}$$

The information content (in bits) of a polymer with n^* or fewer differences from the ideal sequence is denoted I and given by

$$I = -\log_2(\Omega^{-1}) = \log_2\Omega. \tag{A19}$$

Note that if $n^* = 0$ then $\Omega = m^L$ and $I = L\log_2 m$.