BIOLOGY DIRECT

BIOLOGY DIRECT

RESEARCH   **Open Access**

Rate of sequence divergence under constant selection

Alexey S Kondrashov[1], Inna S Povolotskaya[2], Dmitry N Ivankov[2,3], Fyodor A Kondrashov[2*]

Abstract

Background: Divergence of two independently evolving sequences that originated from a common ancestor can be described by two parameters, the asymptotic level of divergence E and the rate r at which this level of divergence is approached. Constant negative selection impedes allele replacements and, therefore, is routinely assumed to decelerate sequence divergence. However, its impact on E and on r has not been formally investigated.

Results: Strong selection that favors only one allele can make E arbitrarily small and r arbitrarily large. In contrast, in the case of 4 possible alleles and equal mutation rates, the lowest value of r, attained when two alleles confer equal fitnesses and the other two are strongly deleterious, is only two times lower than its value under selective neutrality.

Conclusions: Constant selection can strongly constrain the level of sequence divergence, but cannot reduce substantially the rate at which this level is approached. In particular, under any constant selection the divergence of sequences that accumulated one substitution per neutral site since their origin from the common ancestor must already constitute at least one half of the asymptotic divergence at sites under such selection.

Reviewers: This article was reviewed by Drs. Nicolas Galtier, Sergei Maslov, and Nick Grishin.

Background

Constant selection mostly manifests itself as negative selection, because permanently beneficial alleles tend to become common. Negative selection constrains and decelerates evolution. Slower evolution of more functionally important nucleotide sites, such as nonsynonymous coding sites, compared to less important sites, such as synonymous sites, is one of the most pervasive patterns in evolution at the sequence level and reflects ubiquity of negative selection and relative rarity of positive selection. Constant selection can accelerate evolution when it favors more mutable alleles [1], but the magnitude of this effect is substantial only when a number of conditions are met [2].

Still, the impact of constant selection on the rate of evolution, manifested in the dynamics of divergence of two independently evolving sequences that originated from a common ancestor, is not as simple as one might think, and should be viewed from two complementary

perspectives. Indeed, constant selection affects both 1) the asymptotic value of sequence divergence E, reached by two sequences after a long time, and 2) the rate r at which the divergence between two sequences approaches E. Here, we separately study the impact of constant selection on E and on r.

Model

Consider one locus (site) with I alleles $A_1, ... A_I$ with frequencies $x_1, ..., x_I$, and fitnesses $1, 1-s_2, ... 1-s_I$, where s_i are coefficients of selection (without loss of generality, the fitness of one of the alleles can be assumed to be 1). A population is either haploid or diploid with intermediate dominance. The rate of mutation of the ith allele into the jth allele is $\mu_{i, j}$, and the effective size of the population is N_e. We assume that $N_e\mu_{i, j} << 1$ for all i, j, so that most of the time one of the alleles is fixed in the population [3], and the dynamics of the model consists of allele replacements occurring at a site at random moments. We consider dynamics of divergence of two long sequences of sites, all with the same properties, that were identical at the moment when they originated from their last common

* Correspondence: fyodor.kondrashov@crg.es
[2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation, C/Dr Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Spain

ancestor and evolved independently after this moment. The level of divergence D will be characterized by the fraction of homologous sites occupied by different alleles.

Results

Asymptotic value of divergence E
Obviously, after enough time passes since the moment when two sequences begun to diverge, all traces of over-random similarity between them are lost, and D reaches its asymptotic value E, given by

$$E = 1 - \sum_{i=1}^{I} X_i^2 \tag{1}$$

where X_i is the equilibrium value of x_i under given mutation and selection. X_i can be found from a set of I linear equations (eq. 10 from [3]; eq. (11) below presents this set in the case of $I = 4$). A strong enough constant selection favoring one allele over all others can make E arbitrarily close to 0.

The rate r of approaching E: selective neutrality
Let us now concentrate on how D approaches E. We will characterize the rate of this approach by r, the power of an exponential decline of the deviation of D from E, and, equivalently, by the half-approach period $\tau = ln2/r$. Indeed, we will see that D approached E exactly (in special cases) or approximately (in the general case) exponentially.

Let us first consider the case of no selection: $s_i = 0$ for all i. Let us also assume that $\mu_{i,j} = \mu$ for all i, j. Then,

$$\begin{aligned} \frac{dD}{dt} &= -2\mu D + 2\mu(I-1)(1-D) \\ &= 2(I-1)\mu(1 - \frac{I}{I-1}D) \end{aligned} \tag{2}$$

because the probability of a mismatch becoming a match during a short interval of time dt is $2\mu dt$, and the probability of a match becoming a mismatch is $2(I-1)\mu dt$ (see Chapters 3 and 4 from [4]). The solution of this equation with $D = 0$ at time 0 can be obtained from

$$\int_0^D \frac{d\Delta}{1 - \frac{I}{I-1}\Delta} = 2(I-1)\int_0^t d\tau \tag{3}$$

and is given by

$$D_{I,neutr}(t) = \frac{I-1}{I}(1 - e^{-2I\mu t}) \tag{4}$$

Thus, $E = 1/2, 2/3$, and $3/4$; and $r = 4\mu, 6\mu$, and 8μ, with 2, 3, and 4 selectively neutral alleles (Figure 1). We will mostly assume that our loci are nucleotide sites, so that these three cases are particularly relevant. Note,

that nucleotide sites where strong selection acts against one or two "forbidden" nucleotides but fitnesses of the remaining three or two "permitted" nucleotides are the same are described by the model with 3 or 2 selectively neutral alleles.

The rate r of approaching E: two alleles with selection
Let us now consider the case of only two alleles A_1 and A_2, under an arbitrary selection, described by $s_2 = s$. We will keep the assumption that $\mu_{i,j} = \mu$ for all i, j, which in this case simply means that the rates of reciprocal mutations are the same. The dynamics of D are described by (5)

where $x_{i,j}$ is the frequency of pairs of sites occupied by allele A_i in one sequence and allele A_j in the other sequence (due to symmetry of the two sequences, we will use this notation assuming that $i \leq j$), and $f_{i,j}$ is the rate at which allele i is replaced with allele j. Then, $D = x_{1,2}$ and $x_{1,1} = x_1 - 0.5D$, because $x_1 = 0.5x_{1,2} + x_{1,1}$. Analogously, $x_{2,2} = 1 - x_1 - 0.5D$, because $x_2 + x_1 = 1$.

Let us assume that the ancestral sequence from which the two sequences under consideration begun their independent divergence had already attained equilibrium allele composition. Then, $x_1 = X_1$ is invariant and is equal to $e^S/(e^S + 1)$ (eq. 6 from [3]). According to eq. 7 from [3], $f_{1,2} = -\mu S/(1-e^S)$ and $f_{2,1} = \mu S/(1-e^{-S})$, where $S = 4N_e s$, in the case of diploidy. Thus, the equation (5) can be written as:

$$\begin{aligned} \frac{dD}{dt} &= \frac{2\mu S}{1-e^{-S}}\left(\frac{e^{-S}}{1+e^{-S}} - \frac{1}{2}D \right) \\ &\quad - \frac{2\mu S}{1-e^{S}}\left(\frac{e^{S}}{1+e^{S}} - \frac{1}{2}D \right) \\ &\quad - D\mu S\left(\frac{1}{1-e^{-S}} - \frac{1}{1-e^{S}} \right) \end{aligned} \tag{5}$$

with a solution:

$$\begin{aligned} D_{2,any\_selection}(t) = \\ \frac{2e^{S}}{\left(e^{S}+1\right)^2}\left(1 - \exp\left(-2\mu S \frac{e^{S}+1}{e^{S}-1}t \right) \right) \end{aligned} \tag{6}$$

Thus,

$$E = \frac{2e^{S}}{\left(e^{S}+1\right)^2} \tag{7}$$

and

$$r(S) = 2\mu S \frac{e^{S}+1}{e^{S}-1} \tag{8}$$

It is easy to show that $r(S)$ has a single minimum at $S = 0$. Therefore, with only two permitted alleles the lowest rate $r(0) = 4\mu$ of approaching E, and the longest half-approach period $\tau = ln2/(4\mu)$ is achieved under neutrality (4), where the value of $E = 1/2$ is the highest, because $X_1 = X_2$. When S increases, E rapidly approaches zero, and $r(S)$ slowly increases without a limit (Figure 2). A more general case of unequal rates of reciprocal mutations can also be considered using eq. 7 from [3].

The rate r of approaching E: general 4-allele case

Let us now consider the general case with $I = 4$. Here it will be more convenient to use letters A, T, G, and C to denote the four alleles. With 4 alleles under selection, D becomes dynamically insufficient, and the divergence of sequences needs to be described in terms of frequencies of the six possible mismatches $x_{i,j}$ ([5], eq. 1):

$$\frac{dx_{A,T}}{dt} = 2x_{A,A}f_{A,T} + 2x_{T,T}f_{T,A} - x_{A,T}(f_{A,notA} + f_{T,notT})$$
$$+ x_{A,G}f_{G,T} + x_{A,C}f_{C,T} + x_{T,G}f_{G,A} + x_{T,C}f_{C,A}$$

$$\frac{dx_{A,G}}{dt} = 2x_{A,A}f_{A,G} + 2x_{G,G}f_{G,A} - x_{A,G}(f_{A,notA} + f_{G,notG})$$
$$+ x_{A,T}f_{T,G} + x_{A,C}f_{C,G} + x_{T,G}f_{T,A} + x_{G,C}f_{C,A}$$

$$\frac{dx_{A,C}}{dt} = 2x_{A,A}f_{A,C} + 2x_{C,C}f_{C,A} - x_{A,C}(f_{A,notA} + f_{C,notC})$$
$$+ x_{A,T}f_{T,C} + x_{A,G}f_{G,C} + x_{T,C}f_{T,A} + x_{G,C}f_{G,A}$$

$$\frac{dx_{T,G}}{dt} = 2x_{T,T}f_{T,G} + 2x_{G,G}f_{G,T} - x_{T,G}(f_{T,notT} + f_{G,notG})$$
$$+ x_{A,T}f_{A,G} + x_{A,G}f_{A,T} + x_{T,C}f_{C,G} + x_{C,G}f_{C,T}$$

$$\frac{dx_{T,C}}{dt} = 2x_{T,T}f_{T,C} + 2x_{C,C}f_{C,T} - x_{T,C}(f_{T,notT} + f_{C,notC})$$
$$+ x_{A,T}f_{A,C} + x_{A,C}f_{A,T} + x_{T,G}f_{G,C} + x_{G,C}f_{G,T}$$

$$\frac{dx_{G,C}}{dt} = 2x_{G,G}f_{G,C} + 2x_{C,C}f_{C,G} - x_{G,C}(f_{G,notG} + f_{C,notC})$$
$$+ x_{A,G}f_{A,C} + x_{A,C}f_{A,G} + x_{T,G}f_{T,C} + x_{T,C}f_{T,G}$$

$$(9)$$

where the rates of allele replacements $f_{i,j}$, as before are given by eq. 7 from [3] with the coefficient of selection reflecting the difference between fitnesses of the ith and the jth allele and $f_{i,noti}$ is the sum of rates of replacement of ith allele into all other alleles. Frequencies of the four matches are inferred from the dynamical variables: $x_{A,A} = X_A - 0.5(x_{A,T} + x_{A,G} + x_{A,C})$; $x_{T,T} = X_T - 0.5(x_{A,T} + x_{T,G} + x_{T,C})$; $x_{G,G} = X_G - 0.5(x_{A,G} + x_{T,G} + x_{G,C})$; $x_{C,C} = X_C - 0.5(x_{A,C} + x_{T,C} + x_{G,C})$, where X_i, are obtainable from [3] (eq 10)

$$-x_A f_{A,notA} + x_T f_{T,A} + x_G f_{G,A} + x_C f_{C,A} = 0$$
$$-x_T f_{T,notT} + x_A f_{A,T} + x_G f_{G,T} + x_C f_{C,T} = 0$$
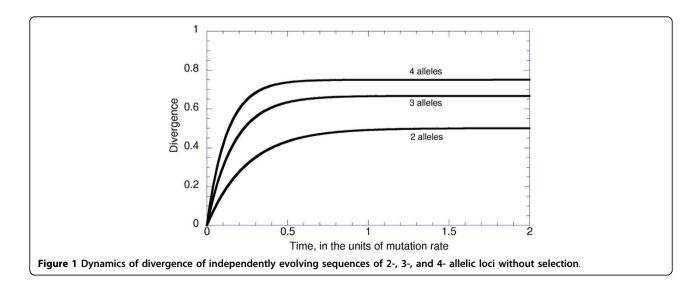$$-x_G f_{G,notG} + x_A f_{A,G} + x_T f_{T,G} + x_C f_{C,G} = 0$$
$$-x_C f_{C,notC} + x_A f_{A,C} + x_T f_{T,C} + x_G f_{G,C} = 0$$

$$(10)$$

under the assumption that $x_A + x_T + x_G + x_C = 1$. D is the sum of all the six variables in (10).

Numerical investigation of (10) revealed the following facts:

1) D approaches E strictly exponentially only in the case of just two permitted alleles. Otherwise the approach is decelerating: the time it takes for D to decline from 0 to $E/2$, $t_{0,1/2}$ is shorter than the time $t_{1/2,3/4}$ during which D declines from $E/2$ to $3E/4$. However, this effect is weak: $t_{1/2,3/4}$ never exceeds $t_{0,1/2}$ by more than 10%. Thus, we can safely regard $t_{0,1/2}$ as the half-approach period τ.

2) D approaches E at a rate that cannot be below $r = 4\mu$, which is only 2 times less than $r = 8\mu$, the rate for the case of 4 neutral alleles ($S_2 = S_3 = S_4 = 0$) (eq 4). This lowest $r = 4\mu$ was observed only in the case of 2 neutral alleles ($S_2 = 0$; $S_3 = S_4 \to 8$) (eq 9). In contrast, r can be arbitrarily large, but this happens only when just one allele is strongly favored ($S_2, S_3, S_4 \to 8$)

Figure 1 Dynamics of divergence of independently evolving sequences of 2-, 3-, and 4- allelic loci without selection.

(eq 9). When selection is moderate ($S \leq 2$), $4\mu < r < 8\mu$ (Figure 3).

3) If the assumption of equal mutation rates is relaxed, the expected value of r under a particular mode of selection does not change much (Figure 4).

Discussion

We have found a contrast between how constant selection affects the asymptotic level of divergence of independently evolving sequences E and the rate r at which this level is approached. While E always declines and can be affected arbitrarily strongly, r can both increase and decrease. The increase of r can be arbitrarily large, and occurs under conditions when E approaches 0, *i. e.* when selection strongly favors one allele over all others. However, r cannot decline by more than a factor of two (assuming equal rates of all mutations) relative to the case of 4 selectively neutral alleles, with the lowest r attained when two alleles have equal high fitnesses and the other alleles are strongly selected against. Unequal mutation rates apparently do not affect this conclusion. Of course, if selection coincidently permits only those two alleles that rarely mutate into each other, it can lead to an arbitrary decrease of r. However, in a general situation, unequal mutation rates do not substantially affect the impact of selection on r.

In other words, the highest, for the case of 4 alleles, asymptotic divergence $E = 3/4$ is approached with a half-approach period $\tau = ln2/(8\mu) \approx 0.087\,\mu^{-1}$, which is neither the shortest nor the longest possible under constant selection. However, the longest possible half-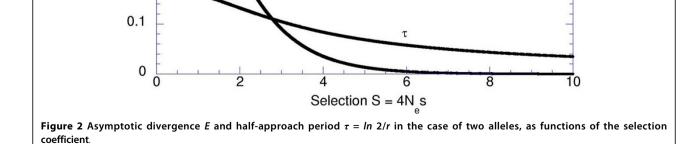approach period $\tau_{max} = ln2/(4\mu) \approx 0.173\,\mu^{-1}$, is just two times longer. Thus, the assertion that constant selection impedes evolution mostly describes its impact on the asymptotic level of divergence, and not on the rate at which this level is attained.
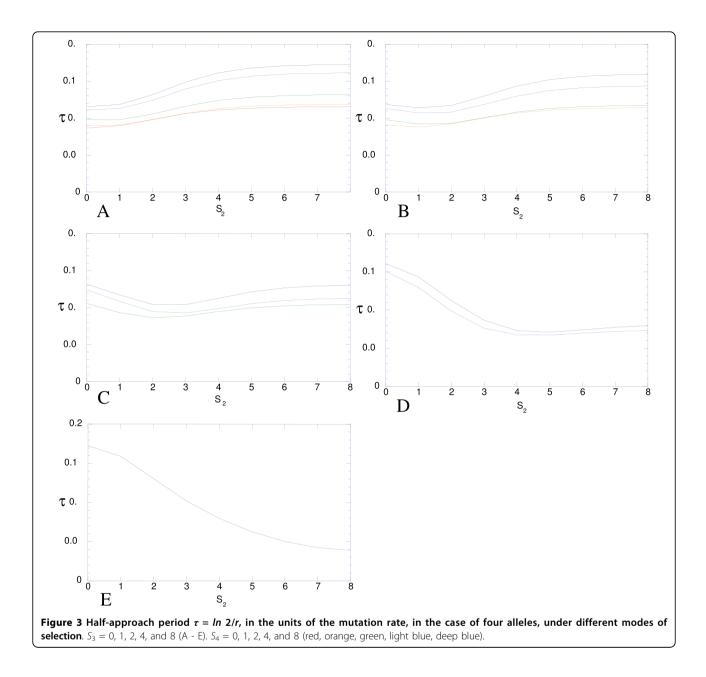
After D travels half of its way to E in the case of 4 neutral alleles, $D = 3/8$. For this D, application of Jukes-Cantor formula [4], (eq. 4.2) produces the following value of the per-site number of nucleotide substitutions per a selectively neutral site:

$$K_{1/2,neutral} = -\frac{3}{4}\ln(1 - \frac{4D}{3}) = \frac{3}{4}\ln(2) \approx 0.52 \qquad (11)$$

Indeed, under selective neutrality the rate of evolution is equal to mutation rate, so that $K_{1/2,\,neutral} = 6\,\mu\tau_{neutral}$ with $I = 4$, and $\tau_{neutral} = ln2/(8\mu)$ in this case. Thus, because constant selection cannot reduce the rate of approaching E by more that a factor of 2, at all kinds of sites under constant selection divergence between two species with $K_{neutral} = 1.04$ must already represent, at the very least, 50% of their equilibrium divergence, as long as the allelic composition of the diverging sequences was at equilibrium from the very beginning. And divergence of all orthologous sequences that are under constant selection in species with $K_{neutral} > 5$ must be already very close to its asymptotic level.

It did not escape our attention that the dynamics of divergence of functionally important sequences are usually inconsistent with this key property of the model considered here: orthologous sequences keep diverging

Figure 2 Asymptotic divergence E and half-approach period $\tau = ln\,2/r$ in the case of two alleles, as functions of the selection coefficient.

Figure 3 Half-approach period $\tau = \ln 2/r$, in the units of the mutation rate, in the case of four alleles, under different modes of selection. $S_3 = 0, 1, 2, 4,$ and 8 (A - E). $S_4 = 0, 1, 2, 4,$ and 8 (red, orange, green, light blue, deep blue).
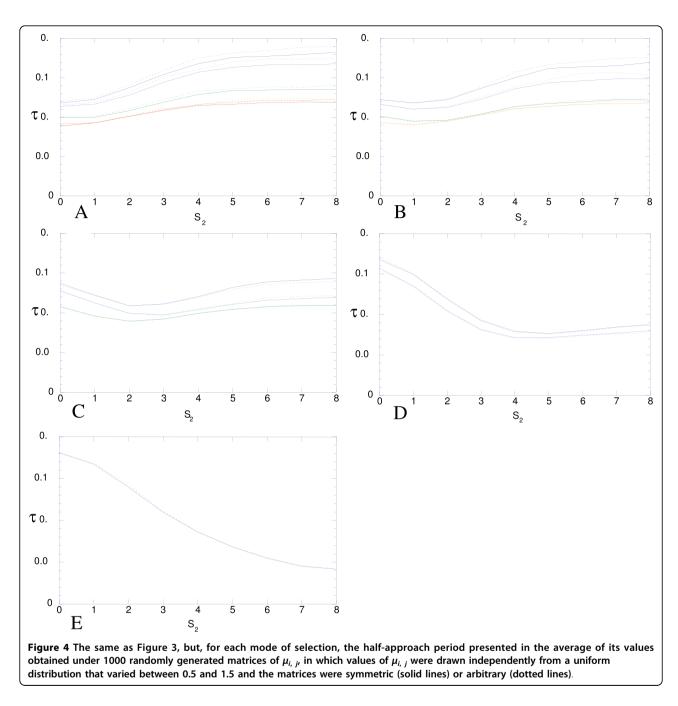
even between very evolutionarily remote genomes. There would be little hope to reconstruct the history of a molecule from extant sequences if selection coefficients were constant in time. In this case, sequences having diverged long enough would essentially include only invariant (strongly selected) and saturated (weakly selected) sites. Since we know that protein sequences can actually help in measuring divergences and building deep phylogenies, this suggests that variations in time of N_e and/or of selection coefficients at individual sites are common, and that their rate of change is what controls in the first place the amount of amino-acid differences between distantly related extant sequences.

Reviewers' comments
Reviewer 1
Referee 1:
Nicolas Galtier
CNRS-Université Montpellier II
Montpellier, France

This manuscript implements a very good idea, namely injecting population genetic parameters, typically used to model microevolution, in substitution matrices, typically used to model long-term sequence evolution. The two must be connected in principle, yet classical Markov models used in phylogenetics are made of arbitrary parameters.

Figure 4 The same as Figure 3, but, for each mode of selection, the half-approach period presented in the average of its values obtained under 1000 randomly generated matrices of $\mu_{i,j}$, in which values of $\mu_{i,j}$ were drawn independently from a uniform distribution that varied between 0.5 and 1.5 and the matrices were symmetric (solid lines) or arbitrary (dotted lines).

These theoretical results are good to know and keep in mind, even though they perhaps do not apply so frequently to real data, as correctly acknowledged, given the assumptions of constant in time population size and selection coefficients. At any rate, this piece might stimulate other studies trying to understand long-term sequence evolution in the light of population genetics. Here are a couple of more specific comments.

1. I am not sure if the model is haploid or diploid. Diploidy is apparently assumed at some stage, but since there is no description of the fitness of a diploid genotype, I guess codominance is assumed. In this case, the diploid model is equivalent to the haploid one, so I would suggest to keep it haploid throughout for simplicity. Or if you make it diploid, please discuss dominance effects.

Indeed, this needed to be clarified. Bulmer's framework we used can handle either haploidy or diploidy with codonimance. We assumed the latter case, but the only difference this made is that $S = 4N_e\ s$, instead of $S = 2N_e\ s$ with haploidy. This is now stated.

2. I feel like a better job could be made in accounting for, and referring to, the 80's/90's literature on Markov models for pairwise distance calculation between sequences. The papers by Rodriguez et al [5] and Lockhart et al [6], for instance, include equations very close to equation 10 of this manuscript. The matrix expressions and calculations used in these papers might help, by the way - I suspect that the numerical exploration of equation 10 performed by the authors could be done analytically. Once you have expressed the substitution rate from allele i to allele j in terms of mutation, selection and drift, you are within Rodriguez's formalism.

Indeed, we needed to cite Rodriguez et al.[5], because our eq. 10 is essentially identical to theirs basic equation 1. However, as Dr. Galtier noted, Rodriguez *et al.* [5]did not study the dependence of solution to this equation on mutation and selection parameters separately; instead they only considered fluxes of allele substitutions. In contrast, we do not see how our results are related to that of Lockhart *et al.*[6], because they were primarily concerned with taking into account different nucleotide compositions of different sequences, and we, as well as Rodriguez *et al.*[6], assumed these to be invariant. Of course, our eq. 10 can, as a system of linear ordinary differential equation, be solved analytically using matrix exponents. However, it will probably still be necessary to investigate this solution numerically, to understand its properties.

3. If I understand correctly, this manuscript demonstrates that there would be little hope to reconstruct the history of a molecule from extant sequences if selection coefficients were constant in time. In this case, sequences having diverged long enough would essentially include only invariant (strongly selected) and saturated (weakly selected) sites. Since we know that protein sequences can actually help in measuring divergences and building deep phylogenies, this suggests that variations in time of Ne and s are common, and that their rate of change is what controls in the first place the amount of amino-acid differences between extant sequences - at least as far as nonadaptive changes are concerned. This conclusion is reminiscent of Gillespie's thoughts about the overdispersed molecular clock. It makes the more or less clock-like behavior of many known molecules even more amazing to me.

We are very thankful for this interpretation of our main conclusion, and we plagiarize it for Discussion with minor changes, placing it at the very end of the paper. In contrast, we do not see a direct correspondence between our results and Gillespie's overdispersed molecular clock. Indeed, changes of selection, which can lead to prolonged divergence, do not necessarily need to occur at random moments of time.

Reviewer 2
Referee 2:
Sergei Maslov
Brookhaven National Laboratory Upton, NY, USA

The conclusions of this study can be summarized in the following statement: the strength of selection has a relatively small effect of the rate at which the evolutionary steady state is approached. Authors offer detailed derivations for the case of a locus consisting of a single nucleotide site. Hence their number of alleles is limited to 4 (A, C, G, and T). In cases where the selection indeed operates at the level of a single nucleotide the derivation presented in the manuscript is through and exhaustive. The mathematically transparent summary of the central results is contained in Eqs. (8) and (9).

I found the discussion part of this manuscript to be somewhat cryptic. For example, I didn't get the point authors were trying to make by using the celebrated Watson and Crick opening sentence in: "It did not escape our attention that the dynamics of divergence of functionally important sequences are usually inconsistent with this key property of the model considered here". What key property of their model authors are talking about?

We expanded this place a bit, using insightful comments of Dr. Galtier. We believe that the main reason why our analysis may be interesting is that it provides a proof by contradiction to the assertion that selection at individual sites keeps constantly changing, which has long been suspected but never demonstrated convincingly. However, this is a separate story.

Another confusing part of the discussion is that after deriving K_1/2,neutral = 0.52 in Eq. (12) the authors state that sequences diverged to the level of

K_neutral = 2K_1/2,neutral = 1.04 represent at least 50% of their equilibrium divergence. I thought that is exactly how K_1/2,neutral was defined. Why double it? In short, I feel the manuscript would benefit from giving a bit more details in the discussion session.

Double it because constant selection can, in the extreme situation, half the rate at which the asymptotic level of divergence is approached. This is now explained.

Another thing, I would like authors to better describe what practical questions their mathematical observations help to understand? I am shooting in the dark here but do these results might explain a puzzlingly uniform rate at which non-neutral substitutions accumulate in the long-term evolutionary experiment by Lenski and collaborators [7]? Having a few concrete examples where the mathematical treatment presented in the manuscript is important would greatly improve the readability of the manuscript.

 One cannot be sure that Lenski's lab data are relevant to our analysis. Indeed, they placed their bacteria in a novel environment, so their system reflects a dynamics of adaptation and may not be at equilibrium, in contrast to our assumptions. We believe that our analysis is important because it demonstrates that, in the long term, evolution of even conservative orthologs does not proceed under constant selection.

Reviewer 3

Referee 3:
Nick Grishin
University of Texas Southwestern Medical Center
Dallas, TX, USA
This reviewer provided no comments for publication.

Author details

[1]Life Sciences Institute, Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA. [2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation, C/Dr Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Spain. [3]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia.

Authors' contributions

FAK designed this study, ASK formulated the theory and performed most of the computational analyses, ISP derived the analytical solution, DNI participated in finding the analytical solution and in performing additional computational analyses. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Eyre-Walker A: **The effect of constraint on the rate of evolution in neutral models with biased mutation.** *Genetics* 1992, **131**:233-234.
2. Kondrashov FA, Ogurtsov AY, Kondrashov AS: **Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites.** *J Theor Biol* 2006, **240**:616-626.
3. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129**:897-907.
4. Li WH: *Molecular Evolution* Sunderland, MA, Sinauer 1997.
5. Rodríguez F, Oliver JL, Marín A, Medina JR: **The general stochastic model of nucleotide substitution.** *Journal of Theoretical Biology* 1990, **142**:485-501.
6. Lockhart PJ, Steel MA, Hendy MD, Penny D: **Recovering evolutionary trees under a more realistic model of sequence evolution.** *Mol Biol Evol* 2004, **11(4)**:605-12.
7. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF: **Genome evolution and adaptation in a long-term experiment with Escherichia coli.** *Nature* 2009, **461**:1243-1247.