Evolutionary stochastic dynamics of speciation and a simple genotype-phenotype map for protein binding DNA

Bhavin S. Khatri and Richard A. Goldstein
MRC National Institute for Medical Research
Mathematical Biology Division
The Ridgeway, London, NW7 1AA, U.K.
(Dated: May 14, 2013)

Speciation is of fundamental importance to understanding the huge diversity of life on Earth. In contrast to current phenomenological models, we develop a biophysically motivated approach to study speciation involving the co-evolution of protein binding DNA for two geographically isolated populations. Our results predict that, despite neutral diffusion of hybrids in trait space, smaller populations have a higher rate of speciation, due to sequence entropy biasing populations more closely to incompatible regions of phenotype space. A key lesson of this work is that non-trivial contributions of sequence entropy give rise to a strong population size dependence on speciation rates.

Speciation is thought to underly much of the diversity of life on Earth today. The development of quantitative models that can predict speciation rates will thus allow better understanding of the different factors that maintain bio-diversity along with the processes of extinction and environmental change [1,2]. Yet the detailed genetic mechanisms by which distinct species arise is still largely not understood. Darwin [3], despite the title of his magnum opus, struggled to understand how natural selection could give rise to hybrid inviability or infertility. If the hybrid inviability were due to a single locus, how could two species evolve from a common ancestor, since at least one of these species would have to evolve past an inviability bottleneck; for example, if a pair of species, which share common ancestry, are fixed for aa and AA alleles, respectively, and the genotype aA is inviable, how did one of these populations evolve the Ao allele? It has since been understood that non-linear or epistatic interactions between different loci can give rise to so-called Dobzhansky-Müller incompatibilities [4–6] between independently evolving lineages. For example, two lines evolving independently through geographic isolation (allopatric evolution) from a common ancestor ab, can fix the allelic combinations aB and Ab respectively, yet the hybrid genotype AB is inviable.

Field data [1,5] suggest that the most dominant form of speciation does indeed involve geographically isolated populations with no or very little gene flow and that the mechanism is commonly via Dobzhansky-Müller incompatibilities [5,6]. There are a number of models of allopatric speciation based on Dobzhansky-Müller incompatibilities, which consider independent lineages evolving neutrally or under varying selection pressures on each line [10–13]. However, all of these models are essentially phenomenological in their assumptions about the genetic basis of speciation. This is understandable without recourse to very detailed models; however, in this paper we develop a generic approach, which explores universal properties that the genetic basis of traits may have on the process of speciation. In particular, much recent work has shown that in general mappings from genotype to phenotype are non-trivial [14–17], giving rise to a number of previously unpredicted effects, due to the increasing prominence that sequence entropy effects have for small populations. In this paper, we examine the process of how incompatibilities arise in allopatry, for an abstract, yet biophysically motivated model of transcription factor binding, which accounts for the sequence entropy of binding and ask what population size effects such a genotype-phenotype maps induces. The binding of transcription factors to DNA to control gene expression is arguably one of the most important co-evolving systems for organisms and crucial for their correct development and so makes an ideal first case study to study the affects of genotype-phenotype maps on speciation.

To tackle this question we develop the tools of stochastic dynamics for evolution in the regime where \( \mu N \ll 1 \), where \( \mu \) is the mutation rate and \( N \) is the population size, and use it to ask how hybrid fitness and the probability of incompatibilities increases with divergence time between a pair of lines under the same stabilising selection pressure. We find, in contrast to current models, which predict no population size effect [10–13], that despite the two lines diverging neutrally, smaller populations have a higher rate of speciation as the effect of sequence entropy is to bias populations more closely to incompatible regions of phenotype space. We suggest a key lesson of this work, is that even though co-evolving loci on different lineages may diverge in a population-size independent neutral manner, non-trivial contributions of sequence entropy in a stabilising fitness landscape give rise to strong population size dependence.
A SMOLUCHOWSKI EQUATION FOR EVOLUTIONARY STOCHASTIC DYNAMICS

Natural selection acts on phenotypes, however, in general, there will not be a one-to-one mapping between phenotypes and genotypes, but instead many genotypes can code for the same phenotype. As has been shown, for small population sizes ($\mu N \ll 1$), where populations are largely monomorphic, genotype-phenotype maps can give rise to a bias in evolution towards phenotypes with larger sequence entropy and in equilibrium is described by a balance between a tendency of phenotypes to increase their fitness and at the same time maximise their sequence entropy [18, 19]. This is embodied by the free fitness, $\Phi(\xi) = F(\xi) + S(\xi)/\nu$, where $F$ is the Malthusian fitness [11] and $S = -(\ln p(\xi))_\xi$ is the sequence entropy, where $p$ is the equilibrium distribution of $\xi$; the vector of traits and $\nu$ is the Lagrange multiplier or effective inverse temperature of the canonical ensemble, which is linear in population size $N$; for the Wright-Fisher process $\nu = 2(N - 1)$ and the Moran process $\nu = N - 1$ [11]. In equilibrium, the free fitness is maximised and the probability distribution of traits is Boltzmann distributed, $p(\xi) = Z^{-1} e^{\Phi(\xi)}$, where $Z = \int d\xi e^{\Phi(\xi)}$.

Out of equilibrium, we expect that stochastic dynamics will give rise to: 1) diffusion in phenotype space with diffusion constant $\mu$, where $\mu$ is the mutation rate of all sites contributing to the phenotype; 2) directed motion driven by gradients in the free fitness function with respect to changes in phenotype $\xi$. The flux of probability in phenotype space will then be given by

$$ J = -\frac{1}{2}\mu \nabla p(\xi) + \frac{1}{\zeta} p(\xi) \nabla \Phi(\xi) \tag{1} $$

where $\zeta$ is a coefficient representing the strength of evolutionary change in response to an evolutionary force (or gradient in free fitness) and the factor of a $\frac{1}{2}$ for the mutation rate comes from converting from a discrete random walk to a continuous one. In equilibrium, detailed balance requires this flux to be zero, from which it is simple to show that

$$ \zeta = \frac{2}{\nu \mu}. \tag{2} $$

which is the evolutionary equivalent of the Einstein relation that relates the friction constant to the diffusion constant of a Brownian particle; here the evolutionary friction constant $\zeta$ is inversely proportional to the mutation rate and population size. We can now express the Smoluchowski Equation in its final form using the continuity equation, $\partial p(\xi)/\partial t = \nabla \cdot (\nabla p(\xi) - \nu p(\xi) \nabla \Phi(\xi))$.

$$ \frac{dp}{dt} = \frac{1}{2} \mu \nabla \cdot (\nabla p(\xi) - \nu p(\xi) \nabla \Phi(\xi)). \tag{3} $$

This Smoluchowski Equation can be converted to an equivalent set of stochastic differential equations [21, 22]

$$ \frac{d\xi_i}{dt} = \frac{1}{2} \nu \mu \frac{\partial \Phi(\xi)}{\partial \xi_i} + \eta_i(t), \tag{4} $$

where $i$ corresponds to the $i^{th}$ trait of $\xi$ and where $\eta_i$ is a white noise Gaussian process with moments $\langle \eta_i(t) \rangle = 0$ & $\langle \eta_i(t) \eta_j(t') \rangle = \delta_{ij} \delta(t - t')$. Eqn 4 is a generalisation of the Ornstein-Uhlenbeck process for phenotypic evolution described in [23, but for an arbitrary free fitness landscape and including the correct population size dependence of the strength of the drift term via the Einstein relation Eqn 2.

A SIMPLE CONTINUOUS MODEL OF TRANSCRIPTION FACTOR BINDING DNA

The two-state approximation [24, 25] for transcription factor (TF) binding assumes that amino acid base pair hydrogen-binding energies are approximately additive and that each nonoptimal interaction increases the energy of binding by the same amount. This suggests replacing DNA and amino acid sequences by binary strings. The binding energy is simply proportional to the Hamming distance $r = \langle g_1 - g_2 \rangle \cdot \langle g_1 - g_2 \rangle$ between a pair of binary sequences, $g_1, g_2$. We can assign a fitness $F_i$ to each value of $r$, either arbitrarily or using some knowledge of the biophysics and function of a particular system. On the other hand, the entropy is given directly by the nature of the binary model; there are many sequences that give the same Hamming distance, precisely $\Omega_r = (2^r)^2 = 2^{2r}$, when $\ell$, the sequence length, is large. So to a good approximation the sequence entropy is quadratic in Hamming distance $r$:

$$ S_r = \frac{2}{\ell} (r - \ell/2)^2 + \text{const.} \tag{5} $$

We see that entropy is maximised for $r = \ell/2$, which corresponds to the Hamming distance that has the largest number of genotypes coding it. In order to take advantage of the effective stochastic dynamics described in the previous section, we replace each sequence $g_i$ with a continuous variable $x_i$ and define a Hamming distance like or binding energy variable as $\xi = |x_1 - x_2|$ [20]. Further, we assume a simple quadratic fitness landscape $F(\xi) = -\frac{1}{2} K F \xi^2$. With sequence entropy given by Eqn. 5 we have (to within a constant) the free fitness given by $\Phi(\xi) = -\frac{1}{2} K F \xi^2 - \frac{1}{2} \ell (\xi - \ell/2)^2$.

Using Eqn. 4 treating $x_1$ and $x_2$ as independent variables, we can write down a pair of stochastic differential equations describing the dynamics of the sequence-like variables in a quadratic landscape:
\[ \frac{dx_1}{dt} = -\frac{1}{2} \nu \mu \kappa (x_1 - x_2) + \mu \text{sgn}(x_1 - x_2) + \eta_1(t), \]
\[ \frac{dx_2}{dt} = -\frac{1}{2} \nu \mu \kappa (x_2 - x_1) + \mu \text{sgn}(x_2 - x_1) + \eta_2(t), \]

where \( \kappa = \kappa_F + \frac{4}{\nu} \), which is the curvature in the free fitness landscape and is a sum of the curvatures due to fitness and sequence entropy. In addition, we have used the fact that \( \partial_z (z - c)^2 = 2(z - c \text{sgn}(z)) \), where \( \text{sgn} \) is the sign function that returns the sign of the argument.

**ANALYTICAL CALCULATION OF HYBRID FITNESS AND THE PROBABILITY OF AN INCOMPATIBILITY**

From Eqn. 8, fitness and sequence entropy balance (maximum in free fitness) for \( x_1 - x_2 = \pm \xi_{\infty} = \pm \frac{2}{\nu} \varphi \), so the free fitness is doubled peaked with a cusp valley at \( x_1 - x_2 = 0 \), except in the limit of an infinite population size (\( \nu \to \infty \)). For the sake of analytical tractability, we can assume that each lineage is always confined to one or the other maximum, such that without loss of generality the initial condition \( x_1(0) > x_2(0) \) remains true. This implies we can make the simplification \( \xi = x_1 - x_2 \), from which it follows:

\[ \frac{dx_1}{dt} = -\lambda (x_1 - x_2) + \mu + \eta_1(t), \]
\[ \frac{dx_2}{dt} = -\lambda (x_2 - x_1) - \mu + \eta_2(t), \]

where the characteristic relaxation rate of the system is given by \( \lambda = \nu \mu \kappa / 2 \). Note that this system of equations is analogous to the dynamics of a pair of overdamped beads connected by spring with a temperature-dependent elastic constant. To solve these equations, it is straightforward to take the Laplace transform of these equations, solve the resulting matrix equation to give solutions in Laplace space and find the inverse Laplace transform to give

\[ x(t) = Jx(0) + \frac{1}{\kappa \nu} (1 - e^{-2\lambda t}) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \int_0^t J(t-t')\eta(t')dt', \]

where \( x = (x_1, x_2)^T \), \( \eta = (\eta_1, \eta_2)^T \), the matrix \( J \) is given by

\[ J = \frac{1}{2} \begin{pmatrix} 1 + e^{-2\lambda t} & 1 - e^{-2\lambda t} \\ 1 - e^{-2\lambda t} & 1 + e^{-2\lambda t} \end{pmatrix}, \]

and it is understood that the integral of the vector above is an element by element operation. There will be an equivalent set of stochastic differential equations for the second lineage of the same form as Eqn. 7 with solution \( x'(t) \) given by Eqn. 8 and binding energy \( \xi' = |x'_1 - x'_2| \) with an identically distributed noise vector, but uncorrelated with the first lineage. If we let \( w = x_1 - x'_2 \) and \( w' = x'_1 - x_2 \) and the hybrid binding energies be \( h = |w| \) and \( h' = |w'| \), then it is straightforward to show using Eqs. 5 and 6 that the moments of the Gaussian process that

\[ \langle w(t) \rangle = \langle w'(t) \rangle = \langle \xi(t) \rangle = \langle \xi'(t) \rangle, \]

\[ \langle w(t) \rangle = \xi_0 e^{-2\lambda t} + \frac{2}{\kappa \nu} (1 - e^{-2\lambda t}). \]

We see that for long times (\( \lambda t \gg 1 \)), the average binding energy \( \langle \xi(\infty) \rangle \) decays from the initial condition to its equilibrium value \( \xi_{\infty} = \frac{2}{\nu} \varphi \) on the timescale \( 1/(2\lambda) \); in reality, using the full dynamics in Eqn. 8 this variable would have zero mean in the long time limit, but as we see below the results are insensitive to this error. If we define the vector \( w = (w, w')^T \), we find that the covariance matrix \( \Sigma = \langle (w - \langle w \rangle)(w - \langle w \rangle)^T \rangle \) is symmetric and has elements

\[ \Sigma_{11} = \mu t + \frac{1}{2\kappa \nu} (1 - e^{-4\lambda t}), \]
\[ \Sigma_{12} = -\mu t + \frac{1}{2\kappa \nu} (1 - e^{-4\lambda t}). \]

We see for short times (\( \lambda t \ll 1 \)), the off-diagonal terms are zero, which suggests the binding energies of the two hybrids are uncorrelated. However, for long times (\( \lambda t \gg 1 \)) they become anti-correlated, which shows, unlike at short times, the probability of both hybrids being DMI or not, is simply, 

A similar calculation of the variance on each lineage gives,

\[ \langle \xi^2 \rangle - \langle \xi \rangle^2 = \frac{1}{\kappa \nu} (1 - e^{-4\lambda t}), \]

which shows that, as expected, the co-evolutionary constraint of the free fitness landscape bounds the variance of binding energies on each lineage. So the variance in the hybrid binding energies is due to a pure diffusive term \( \mu t \), which represents how the two lineages diffuse apart by independent mutations, plus a term which represents the saturating growth of variance of each lineage after divergence.

The dynamics of the probability of a DMI for each hybrid, irrespective of whether the other hybrid has a DMI or not, is simply,

\[ P_j(t) = 1 - \int_{-\xi^*}^{\xi^*} p(w, t)dw, \]
where $\xi^* = \sqrt{2|F^*|/\kappa_F}$, where $F^*$ is the threshold fitness below which an incompatibility arises in a hybrid. The variable $w$ is given by the sum of a number of Gaussian processes, so $p(w, t)$ itself must be Gaussian (with $p(h, t) = p(w, t) + p(-w, t)$), which is completely specified by its mean (Eqn 10) and variance $\Sigma_{11} = \langle w^2 \rangle - \langle w \rangle^2$, which can be directly calculated from Eqns 10 & 11.

From Eqn 14 the probability of a DMI is then simply an integral of a Gaussian, which is expressed in terms of complementary error functions:

$$P_I(t) = \frac{1}{2} \text{erfc} \left( \frac{\xi^* - \langle w \rangle}{\sqrt{2\Sigma_{11}}} \right) + \frac{1}{2} \text{erfc} \left( \frac{\xi^* + \langle w \rangle}{\sqrt{2\Sigma_{11}}} \right).$$

Note that both the average hybrid fitness and the probability of incompatibilities (Eqn 14) collapse to the same curves as a function of $\mu t$, for the same value of $\kappa \nu = \kappa_F \nu + 4/\ell$ and when fitness is measured relative to the curvature $F/\kappa_F$.

**RESULTS**

We can compare the analytical calculations of the previous section with numerical simulations of Eqns. 6 for a pair of lines diverging from a common ancestor, where no approximation is made regarding the values of $x_1$ and $x_2$. The results of the simulations are averaged over $10^4$ independent realisations and all results assume an effective sequence length $\ell = 10$, $F^*/\kappa_F = -25$ and $\xi(0) = \xi_\infty = 2/\kappa \nu$, which is the mean value of $\xi$ in equilibrium.

In Fig 1, we have plotted the probability of an incompatibility, for various values of $\kappa \nu$, where solid lines are the analytical calculation and the dotted lines are from the numerical integration of Eqn 6. Firstly, we see that the analytical predictions compare very well to integrating the full stochastic differential equations, validating our simplifying assumptions. Secondly, we see there is a large population size effect for the probability of an incompatibility; the characteristic time for incompatibilities to arise is shorter as the population size decreases, where in addition for small populations ($\kappa_F \nu \ll 0.1$) there are two such characteristic timescales. In addition, we see that the dynamics of $P_I(t)$ become insensitive to changes in population size both for small population sizes ($\kappa_F \nu \ll 0.1$) and large population sizes ($\kappa_F \nu \gg 0.1$). Note that, as we have not bounded the binding energies to a maximum value of $\ell$, we would expect real sequences to have a smaller plateau value of $P_I$ in the long time limit of $\mu t \gg 1$; as we argue below it is the short time limit that is most relevant to speciation.

Under this general behaviour, we can consider what happens in the phase space of $x_1$ and $x_2$ for the first lineage, versus the phase-space for $x_1$ and $x_2'$ for the hybrids, as shown in Fig 2. Initially, both lineage and hybrid populations diffuse in a neutral and spherically symmetric manner (like $2\mu t$ and $4\mu t$, respectively) up to the time $\sim \lambda^{-1}$, when the change in free fitness is of order the mean fitness $\sim 1/\nu$ and the accumulated variance of the hybrid population approaches the characteristic width of the potential $\delta \xi^2 \sim 1/\kappa \nu$. After this time the co-evolutionary constraint of the free fitness landscape is felt on each lineage and the probability density is then squeezed along a tube whose axis is defined by $x_1 = x_2 = x_\infty$ (assuming an initial condition $x_1(0) > x_2(0)$ and width $\delta x$. As the marginal pdfs for $x_1$ and $x_2'$ will be identical, the result is that in the hybrid phase-space, we have that the $p(x_1, x_2', t)$ still grows in a spherically symmetric manner, as indicated by the term $2\mu t$ in Eqn 11 incompatibilities arise when hybrid populations have diffused to one or the other critical binding energy at $x_1 - x_2' = \pm \xi^*$. From Eqn 14 and Fig 2, we see that there will in general be two characteristic times for DMIs to arise, given by the condition that $(\xi^* \pm 2/\kappa \nu)^2 \sim \Sigma_{11}(t)$. It is then simple to see that in the limit of $\kappa_F \nu \gg 4/\ell$, $\xi_0 \to 0$ and the time to diffuse to each boundary is the same as observed in Fig 1. Finally, for small population sizes ($\kappa_F \nu \ll 4/\ell$) $\kappa \nu \to 4/\ell$, which shows the dynamics of DMIs becomes independent of population size in this limit as well. It is clear that the population size dependence described here is likely to be also seen in more complex models of co-evolving loci as for a general free fitness landscape, the balance between sequence entropy and fitness is population size dependent, poising populations nearer or further away from such incompatible regions.
For the actual process of speciation, it will typically be the short-time behaviour that will dominate. Given a large number $M \sim 1000 \rightarrow 10000$ of similar non-interacting clusters of loci, we would expect speciation when $MP_1 \gg 1$ and so the critical probability of a DMI will be $\sim 1/M \sim 10^{-4} \rightarrow 10^{-3}$. As shown in the inset of Fig. 1 which is a plot of $P_1(t)$ on a log-log scale together with the threshold probability $1/M = 10^{-4}$, there is strong population size dependence on the time to speciation. In particular, we see this time is relatively insensitive to the exact value of $M$, compared to the population size effect we describe and arises due to the steep gradient of $P_1(t)$ at short times.

The process of speciation underlies the vast diversity of life on Earth. However, conventional models of speciation via Dobzhansky-Müller incompatibilities for geographically isolated populations with neutral divergence predict that speciation is independent of population size \cite{10-13}. However, such models are phenomenological with respect to the genetic basis of incompatibilities. Here using the tools of stochastic dynamics and a simple biophysically motivated model for the genotype-phenotype map of protein binding DNA, we have shown that there is a significant population size dependence of the time for hybrid incompatibilities to arise for independently evolving isolated populations under the same stabilising selection pressure. Protein binding DNA to control gene expression is a prototypical co-evolving system and critical for the proper development of organisms, thus these results have strong implications for the speciation rates and diversification of populations at small population sizes. We should note that our model does not account for differential selection across lineages, although it can easily be extended to do so, which can reinforce or have an opposite dependence on population size dependent on whether selection is beneficial or mildly deleterious \cite{13}; the work here highlights a previously unanticipated population size dependence, which is necessary for a fully quantitative model of speciation.

We acknowledge useful discussions with David Pollock, University of Colorado and funding from the Medical Research Council, U.K.