Rapid speciation despite conservation of phenotype

Joshua S. Schiffman[†] Peter L. Ralph[†‡]
[†]University of Southern California, Los Angeles, California [‡]University of Oregon, Eugene, Oregon
jsschiff@usc.edu plr@uoregon.edu

Abstract

It is known that even if a species' phenotype has remained unchanged over evolutionary time, the underlying mechanism may have changed, since distinct molecular pathways can realize identical phenotypes. In this paper, we use quantitative genetics and linear systems theory to study how a gene network underlying a conserved phenotype evolves as genetic drift of small mutational tweaks to the molecular pathways cause a population to explore the set of mechanisms with identical phenotypes. In this setting we treat organisms as "black boxes" for which the environment provides input and the phenotype is the output, and there exists an exact characterization of the set of all mechanisms that give the same input–output relationship. We show that in this situation, there is never a unique architecture for any phenotype and that the evolutionary exploration of these distinct and mutationally connected mechanisms can lead to the rapid accumulation of hybrid incompatibilities between allopatric populations. We estimate that in reasonably numerous species, this process could thus the formation of new species over a number of generations proportional to the effective population size.

Additional ideas to consider adding:

- speciation literature
- discuss linearity, linearization, and canalization in introduction
- "On the origin of species not by means of natural selection"

Introduction

A complex molecular machinery translates an organism's genome into the characteristics on which natural selection acts, her phenotype. Attaining a general understanding of the functioning and evolution of this molecular machinery is an overarching goal of many subdisciplines of biology. For example, there is a growing body of data on evolutionary histories and molecular characterizations of particular gene regulatory networks [Jaeger, 2011, Davidson and Erwin, 2006, Israel et al., 2016], as well as thoughtful verbal and conceptual models [True and Haag, 2001, Pavlicev and Wagner, 2012, Weiss and Fullerton, 2000, Edelman and Gally, 2001]. Mathematical models of both particular regulatory networks and the evolution of such systems in general can provide guidance where intuition fails, and thus has the potential to discover general principles in the organization of biological systems and provide concrete numerical predictions [Servedio et al., 2014].

The dynamics of the molecular machinery and its interactions with the environment can be mathematically described in various ways as a dynamical system [Jaeger et al., 2015]. Movement in this direction is ongoing, as researchers have begun to study the evolution of both abstract [Wagner, 1994, 1996, Siegal and Bergman, 2002, Bergman and Siegal, 2003, Draghi and Whitlock, 2015] and empirically inspired computational and mathematical models of gene regulatory networks, [e.g. Mjolsness et al., 1991, Jaeger et al., 2004, Kozlov et al., 2012, 2015, 2014, Crombach et al., 2016, Wotton et al., 2015, Chertkova et al., 2017]. It is well known that in many contexts mathematical models can fundamentally be *nonidentifiable* and/or *indistinguishable* – meaning that there can be uncertainty about an inferred model's parameters or even its claims about causal structure, even with access to complete and perfect data [Bellman and Åström, 1970, Grewal and Glover, 1976, Walter et al., 1984]. Models with different parameter schemes, or even different mechanics can equally accurately predict the observed behavior of a given physical system, but still not actually reflect the internal dynamics of the system. In control theory, where electrical circuits and mechanical systems are often the focus, it is understood that there can be an infinite number of "realizations", or ways to reverse engineer the dynamics of a "black box", even if all possible input and output experiments on the

"black box" are performed [Kalman, 1963, Anderson et al., 1966, Zadeh and Deoser, 1976]. The fundamental nonidentifiability of chemical reaction networks is sometimes referred to as "the fundamental dogma of chemical kinetics" [Craciun and Pantea, 2008]. In computer science, this is framed as the relationship among processes that simulate one another [Van der Schaft, 2004]. Finally, the field of *inverse problems* studies those cases where, even if a one-to-one mapping between model and behavior is possible in theory, even tiny amounts of noise can make inference problems nonidentifiable in practice.

Although nonidentifiability may frustrate the occasional engineer or scientist, viewed from another angle, these concepts can provide a starting point for thinking about externally equivalent systems – systems that evolution can explore, so long as the parameters and structures can be realized biologically. These functional symmetries manifest in convergent and parallel evolution, as well as *developmental systems drift*: the observation that macroscopically identical phenotypes in even very closely related species can in fact be divergent at the molecular and sequence level [True and Haag, 2001, Tsong et al., 2006, Hare et al., 2008, **?**, Stergachis et al., 2014].

EDIT IN THESE REFS? The literature is filled with detailed observations of molecular systems and their diversity. *Diversity doesn't imply systems drift – only if the diverse systems are homologous.* There are examples of significant diversity in the networks underlying processes such as circadian rhythm [Sancar, 2008], cell cycle control [Cross et al., 2011, Kearsey and Cotterill, 2003], pattern formation *cite?*, and metabolism [Lavoie et al., 2009, Martchenko et al., 2007, Dalal et al., 2016, Christensen et al., 2011, Hartl et al., 2007, Alam and Kaminskyj, 2013].

In this paper we outline a theoretical framework to study the evolution of biological systems that can be described by systems of differential equations, such as gene regulatory networks. We study the scenario where the optimal phenotype remains constant over evolutionary time, so that despite strong stabilizing selection for phenotype, neutral drift in the underlying genotype is possible. Results from systems theory provide an analytical description of the set of all linear gene network architectures that yield identical phenotypes, which gives concrete expectations for how any such system can, in principal, undergo systems drift and rewire. Even with stabilization selection, a population will explore the set of all possible phenotypically equivalent gene networks. Consequentially, two populations isolated for a sufficiently long period of time, will likely produce inviable hybrids, despite the absence of adaptation, directional selection, or environmental change.

Methods

We study an abstract model of temporal dynamics of the concentrations of a collection of n coregulating molecules within an organism, that may also be affected by temporally varying signals from the environment. We write $\kappa(t)$ for the vector of n molecular concentrations at time t. There are also m "inputs" determined exogenously to the system, denoted $u(t)$, and ℓ "outputs", denoted $\phi(t)$. The output is merely a linear function of the internal state: $\phi_i(t) = \sum_j C_{ij}\kappa_i(t)$ for some matrix C. Since ϕ is what natural selection acts on, we refer to it as the *phenotype* (meaning the "visible" aspects of the organism), and in contrast refer to κ as the *kryptotype*, as it is "hidden" from direct selection. Although ϕ may depend on all entries of κ, it is usually of lower dimension than κ, and we tend to think of it as the subset of molecules relevant for survival. The dynamics are determined by the matrix of regulatory coefficients, A; a time-varying vector of inputs $u(t)$, and a matrix B that encodes the effect of each entry of u on the elements of the kryptotype. The rate at which the i^{th} concentration changes is a weighted sum of the concentrations of the other concentrations as well as the input:

$$\begin{aligned} \dot{\kappa}(t) &= A\kappa(t) + Bu(t) \\ \phi(t) &= C\kappa(t). \end{aligned} \tag{1}$$

Furthermore, we always assume that $\kappa(0) = 0$, so that the kryptotype measures deviations from initial concentrations. Here A can be any $n \times n$ matrix, B a $n \times m$, and C any $\ell \times n$ dimensional matrix, with usually ℓ and m less than n. We think of the system as the triple $\mathcal{S} = (A, B, C)$, which translates (time-varying) m-dimensional input $u(t)$ into the ℓ-dimensional output $\phi(t)$. Under quite general assumptions, we

can write the phenotype as

$$\phi(t) = Ce^{At}\kappa(0) + \int_0^t Ce^{A(t-s)}Bu(s)ds, \tag{2}$$

which is a convolution of the input $u(t)$ with the system's *impulse response*, which we denote as $h(t) := Ce^{At}B$.

Although many different biological systems can be modeled with this approach, for clarity, we focus on gene regulatory networks. In this interpretation, A_{ij} determines how the j^{th} transcription factor regulates the i^{th} transcription factor. If $A_{ij} > 0$, then κ_j upregulates κ_i, while if $A_{ij} < 0$, then κ_j downregulates κ_i. The ith row of A is therefore determined by genetic features such as the strength of j-binding sites in the promoter of gene i, factors affecting chromatin accessibility near gene i, or basal transcription machinery activity. The form of B determines how the environment influences transcription factor expression levels, and C might be the rate of production of a downstream enzyme (although other arrangements could be made).

Here we have assumed that the system is linear, and begins from the "zero" state ($\kappa(0) = 0$). Of course, neither of these are necessarily true for real systems, but the dynamics of most nonlinear systems can be approximated locally by a linear systems near most points. Furthermore, the ease of analyzing linear systems makes this an attractive place to start.

Example 1 (An oscillator)**.** *For illustration, we consider an extremely simplified model of oscillating gene transcription, as for instance is found in cell cycle control or the circadian rhythm. Suppose there are two genes, whose transcript concentrations are given by $\kappa_1(t)$ and $\kappa_2(t)$, and that gene-2 upregulates gene-1 and that gene-1 downregulates gene-2 with equal strength. Furthermore, suppose that only the dynamics of gene-1 are consequential to the oscillator (perhaps the amount of gene-1 activates another downstream gene network). Lastly suppose that the production of both genes is equally upregulated by an exogenous signal. The dynamics of the system are described by*

$$\begin{aligned}\dot{\kappa_1}(t) &= \kappa_2(t) + u(t)\\ \dot{\kappa_2}(t) &= -\kappa_1(t) + u(t)\\ \phi(t) &= \kappa_1(t).\end{aligned}$$

In matrix form the system regulatory coefficients are given as, $A = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$, and $C = \left[\begin{smallmatrix} 1 & 0 \end{smallmatrix}\right]$.

Suppose the input is an impulse at time zero (a delta function), and so its phenotype is equal to its impulse response,

$$\phi(t) = h(t) = \sin t + \cos t.$$

The system and its dynamics are referred to in Figure 1. We return to the evolution of such a system below.

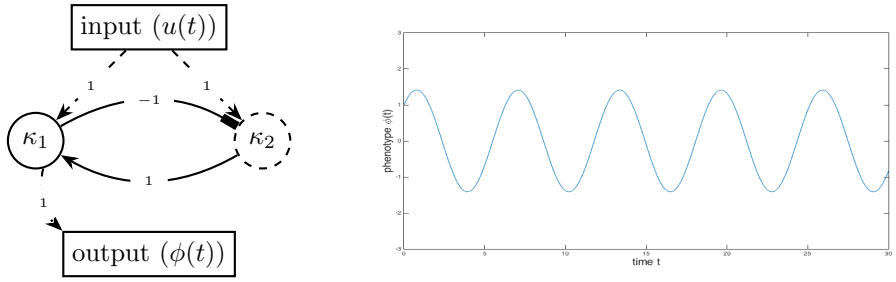

Figure 1: (Left) Graphical representation of the cell cycle control gene network, and (right) plot of the phenotype $\phi(t)$ against time t.

Equivalent gene networks

As reviewed above, some systems with identical phenotypes are known to differ, sometimes substantially, at the molecular level; systems with identical phenotypes do not necessarily have identical kryptotypes. How many different mechanisms perform the same function?

Two systems are equivalent if they produce the same phenotype given the same input, i.e., have the same input–output relationship. We say that the systems defined by (A, B, C) and $(\bar{A}, \bar{B}, \bar{C})$ are **phenotypically equivalent** if their impulse response functions are the same: $h(t) = \bar{h}(t)$ for all $t \geq 0$. This implies that for any acceptable input $u(t)$, if $(\kappa_u(t), \phi_u(t))$ and $(\bar{\kappa}_u(t), \bar{\phi}_u(t))$ are the solutions to equation (1) of these two systems, respectively, then

$$\phi_u(t) = \bar{\phi}_u(t) \qquad \text{for all } t \geq 0.$$

One way to find other systems phenotypically equivalent to a given one is by change of coordinates: if V is an invertible matrix, then the systems (A, B, C) and (VAV^{-1}, VB, CV^{-1}) are phenotypically equivalent because their impulse response functions are equal:

$$\begin{aligned} h(t) &= Ce^{At}B = CV^{-1}Ve^{At}V^{-1}VB \\ &= CV^{-1}e^{VAV^{-1}t}VB = \bar{C}e^{\bar{A}t}\bar{B} = \bar{h}(t). \end{aligned} \tag{3}$$

However, not all phenotypically equivalent systems are of this form: systems can have identical impulse responses without being coordinate changes of each other. In fact, systems with identical impulse responses can involve interactions between different numbers of molecules, and thus have kryptotypes in different dimensions altogether.

This implies that most systems have at least n^2 degrees of freedom, where recall n is the number of components of the kryptotype vector. This is because for an arbitrary $n \times n$ matrix Z, taking V to be the identity matrix plus a small perturbation in the direction of Z above implies that moving A in the direction of $ZA - AZ$ while also moving B in the direction of ZB and C in the direction of $-CZ$ will leave the phenotype unchanged. If A is invertible, for instance, then any such Z will result in a different system.

It turns out that in general, there are more degrees of freedom, except if the system is *minimal* – meaning, informally, that it uses the smallest possible number of components to achieve the desired dynamics. Results in system theory show that any system can be realized in a particular minimal dimension (the dimension of the kryptotype, $n_{\min}$), and that any two phenotypically equivalent systems of dimension $n_{\min}$ are related by a change of coordinates.

Some gene networks, however, can grown or shrink, perhaps following gene duplications and deletions, and also still preserve their phenotypes. More generally, even if the system is not minimal, results from systems theory explicitly describe the set of all phenotypically equivalent systems. We refer to $\mathcal{N}(A_0, B_0, C_0)$ as the set of all systems phenotypically equivalent to the system defined by (A_0, B_0, C_0). Concretely, this is

$$\mathcal{N}(A_0, B_0, C_0) = \left\{(A, B, C) : Ce^{At}B = C_0e^{A_0t}B_0 \text{ for } t \geq 0\right\}. \tag{4}$$

These systems need not have the same kryptotypic dimension n, but must have the same input and output dimensions (ℓ and m, respectively).

The Kalman decomposition, which we now describe informally, elegantly characterizes this set [Kalman, 1963, Kalman et al., 1969, Anderson et al., 1966]. To motivate this, first note that the input $u(t)$ only directly pushes the system in certain directions (those lying in the span of the columns of B). As a result, different combinations of input can move the system in any direction that lies in what is known as the *reachable subspace*. Analogously, we can only observe motion of the system in certain directions (those lying in the span of the columns of C), and so can only infer motion in what is known as the *observable subspace*. The Kalman decomposition then classifies each direction in kryptotype space as either reachable or unreachable, and as either observable or unobservable. Only the components that are both reachable and observable determine the system's phenotype – that is, components that respond to an input and components that produce an observable output.

Concretely, the **Kalman decomposition** of a system (A, B, C) gives a change of basis P such that the transformed system (PAP^{-1}, PB, CP^{-1}) has the following form:

$$PAP^{-1} = \begin{bmatrix} A_{r\bar{o}} & A_{r\bar{o},ro} & A_{r\bar{o},\bar{r}\bar{o}} & A_{r\bar{o},\bar{r}o} \\ 0 & A_{ro} & 0 & A_{ro,\bar{r}o} \\ 0 & 0 & A_{\bar{r}\bar{o}} & A_{\bar{r}\bar{o},\bar{r}o} \\ 0 & 0 & 0 & A_{\bar{r}o} \end{bmatrix},$$

and

$$PB = \begin{bmatrix} B_{r\bar{o}} \\ B_{ro} \\ 0 \\ 0 \end{bmatrix} \qquad (CP^{-1})^T = \begin{bmatrix} 0 \\ C_{ro}^T \\ 0 \\ C_{\bar{r}o}^T \end{bmatrix}.$$

The impulse response of the system is given by

$$h(t) = C_{ro} e^{A_{ro} t} B_{ro},$$

and therefore, the system is phenotypically equivalent to the *minimal* system (A_{ro}, B_{ro}, C_{ro}). (Here the subscript ro refers to the both *reachable and observable* subspace, while $\bar{r}\bar{o}$ refers to the *unreachable and unobservable* subspace, and similarly for $\bar{r}o$ and $r\bar{o}$.)

Any two minimal systems are related by a change of coordinates, and so the minimal subsystems obtained by the Kalman decomposition are unique up to a change of coordinates. In particular, this implies that there is no equivalent system with a smaller number of kryptotypic dimensions than the dimension of the minimal system. It is also remarkable to note that the gene regulatory network architecture to achieve a given input–output map is never unique – both the change of basis used to obtain the decomposition and, once in this form, all submatrices other than A_{ro}, B_{ro}, and C_{ro} can be changed without affecting the phenotype, and so represent degrees of freedom.

Note on implementation: The *reachable subspace*, which we denote by $\mathcal{R}$, is defined to be the closure of span(B) under applying A, and the *unobservable subspace*, denoted $\bar{\mathcal{O}}$, is the largest A-invariant subspace contained in the null space of C. The four subspaces, $r\bar{o}$, ro, $\bar{r}\bar{o}$, and nro are defined from these by intersections and orthogonal complements.

For the remainder of the paper, we interpret $\mathcal{N}$ as the phenotypically neutral landscape, wherein a large population will drift under environmental and selective stasis. Even if the phenotype is constrained and remains constant through evolutionary time, the molecular mechanism underpinning it is not constrained and likely will not be conserved.

Finally, note that if B and C are held constant – i.e., if the relationships between environment, kryptotype, and phenotype do not change – there are *still* usually degrees of freedom. These correspond to distinct genetic networks that perform indistinguishable functions. The following example 2 gives the set of minimal systems equivalent to the oscillator of example 1, that all share common B and C matrices.

Example 2 (All Phenotypically Equivalent Oscillators)**.** *The oscillator of example 1 is minimal, and so any equivalent system is a change of coordinates by an invertible matrix V. If we further require B and C to be invariant then we need $VB = B$ and $CV = C$. Solving these equations, we find that a one-parameter family $(A(\tau), B, C)$ describes the set of all two-gene systems phenotypically equivalent to the oscillator, where*

$$A(\tau) = \frac{1}{\tau - 1} \begin{bmatrix} \tau & -1 \\ 2\tau(\tau - 1) + 1 & -\tau \end{bmatrix} \text{ for } \tau \neq 1.$$

The resulting set of systems, and their dynamics, are depicted in Figure 2

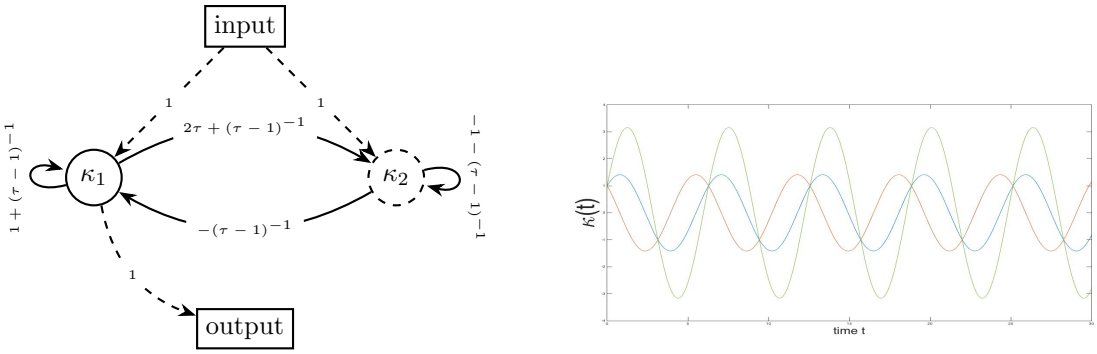

Figure 2: (Left) $A(\tau)$, the set of all phenotype-equivalent cell cycle control networks. (Right) Gene-1 dynamics (blue) for both systems $A(0)$ and $A(2)$ are identical, however, $A(0)$ gene-2 dynamics (orange) differ from $A(2)$ (green).

Sexual reproduction and recombination Parents with phenotypically equivalent, yet different, gene networks may produce offspring with dramatically different phenotypes. If the phenotypes are significantly divergent then the offspring may be inviable or otherwise have very low fitnesses, despite both parents being well adapted. If this is consistent for the entire population, we would consider them to be separate species.

Diploid organisms have two copies of the genome, each of which encodes a set of system coefficients. We assume that a diploid who has inherited systems (A', B', C') and (A'', B'', C'') from her mother and father, respectively, has phenotype determined by the system that averages these two, $((A' + A'')/2, (B' + B'')/2, (C' + C'')/2)$.

Each of the copies of the genome an organism inherits from her parents are generated from their two copies of the genome by meiosis, in which the diploid parent recombines her two genomes. We will assume that each coefficient (i.e., entry of A, B or C) is determined by a single nonrecombining locus, so that each coefficient in the system produced by meiosis is an independent random choice of the two parental coefficients. With these definitions, an F_1 offspring of two individuals carries one system copy from each parent, and an F_2 is the offspring of two independently formed F_1 individuals. If the parents are from distinct populations, these are first- and second-generation hybrids, respectively.

This is a simplification: since the i^{th} row of A summarizes how each gene regulates gene i, and hence is determined by the promoter region of gene i, we would actually expect the elements of a row of A to tend to be inherited together. Similarly, we expect in practice heritable variation in each coefficient to be determined by more than one locus – but this may be a reasonable approximation.

The offspring of two systems with the same phenotype may not have the same phenotype as the parents – in other words $\mathcal{N}$, the set of all systems phenotypically equivalent to a given one, is not, in general, closed under averaging or recombination. Next we discuss how this fact can contribute to hybrid incompatibility and genetic load. If sexual recombination among systems drawn from $\mathcal{N}$ yields systems with divergent phenotypes, populations containing significant diversity in $\mathcal{N}$ can carry genetic load, and isolated populations may fail to produce hybrids with viable phenotypes.

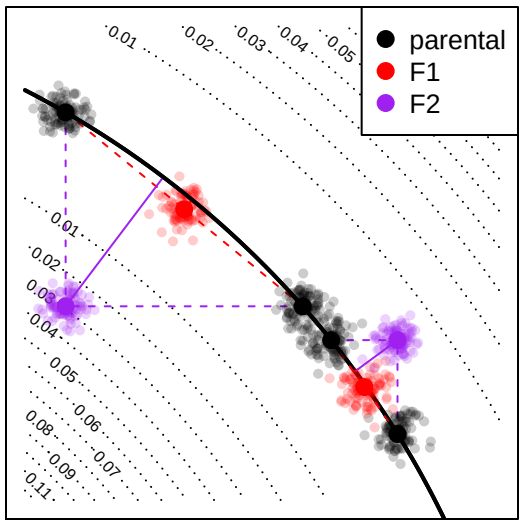

Figure 3: A conceptual figure of the fitness consequences of hybridization: axes represent system coefficients (i.e., entries of A); the line of optimal system coefficients is down in black; dotted lines give phenotypic distances to the optimum. Two pairs of parental populations are shown in black, along the optimum; a hypothetical population of F_1s are shown for each in red, and the distribution of one type of F_2 is shown in purple (other types of F_2 are not shown). Solid lines depict the distance of the F_2 to optimum.

Hybrid incompatibility Two parents with the optimal phenotype can produce offspring whose phenotype is suboptimal if the parents have different underlying systems. This leads to the question: How quickly does the hybrid phenotype break down with increasing distance between the parents? To quantify this, we will measure how far a system's phenotype is from optimal using a weighted difference between impulse response functions. Suppose that $\rho(t)$ is a nonnegative, smooth, square-integrable weighting function, suppose that $h_0(t)$ is the *optimal* impulse response function and define the "distance to optimum" of another impulse response function to be

$$D(h) = \left(\int_0^\infty \rho(t) \|h(t) - h_0(t)\|^2 dt \right)^{1/2}. \tag{5}$$

Consider reproduction between a parent with system (A, B, C) and another displaced by distance ϵ in the direction (X, Y, Z), i.e., having system $(A + \epsilon X, B + \epsilon Y, C + \epsilon Z)$. We assume both are "perfectly adapted" systems, i.e., having impulse response function $h_0(t)$, and their offspring has impulse response function $h_\epsilon(t)$. A Taylor expansion of $D(h_\epsilon)$ in ϵ is explicitly worked out in Appendix F, and shows that the phenotype of an F_1 hybrid between these two is at distance proportional to ϵ^2 from optimal, while F_2 hybrids are at distance proportional to ϵ. This is because an F_1 hybrid has one copy of each parental system, and therefore lies directly between the parental systems (see Figure 3) – the parents both lie in $\mathcal{N}$, which is the valley defined by D, and so their midpoint only differs from optimal due to curvature of $\mathcal{N}$. In contrast, an F_2 hybrid may be homozygous for one parental type in some coefficients and homozygous for the other parental type in others; this means that each coefficient of an F_2 may be equal to either one of the parents, or intermediate between the two; this means that possible F_2 systems may be as far from the optimal set, $\mathcal{N}$, as the distance between the parents. The precise rate at which the phenotype of a hybrid diverges depends on the geometry – in Figure 3, this is depicted as the angle of the black line (the optimal set) with respect to the coordinates.

Example 3 (Hybrid Incompatibility in the Oscillator)**.** *Offspring of two equivalent systems from Example 2 can easily fail to oscillate. For instance, the F_1 offspring between homozygous parents at $\tau = 0$ and*

$\tau = 2$ has phenotype $\phi_{F_1}(t) = e^t$, rather than $\phi(t) = \sin t + \cos t$. However, the coefficients of these two parental systems differ substantially, probably more than would be realistically observed between diverging populations. In figure 4 we compare the phenotypes for F_1 and F_2 hybrids between more similar parents, and see increasingly divergent phenotypes as the difference between the parental systems increases. (In this example, the coefficients of $A(2+\epsilon)$ differ from those of $A(2)$ by an average factor of $1+\epsilon/2$; such small differences could plausibly be caused by changes to promoter sequences.) This divergence is quantified in Figure 5, which shows that mean distance to optimum phenotype of the F_1 and F_2 hybrid offspring between $A(2)$ and $A(2+\epsilon)$ increases with ϵ^2 and ϵ, respectively.

The coefficients in A – i.e., the regulatory coefficients – differ between parents by only a few percent (around 0.5% for $\tau = 2.01$ and 5% for $\tau = 2.1$). This is well within the amount of regulatory coefficient variation we expect to find segregating within real populations (discussed further below). For these small values of ϵ, hybrid phenotypes remain relatively stable, consistent with the idea that natural selection will allow such intrapopulation variation.

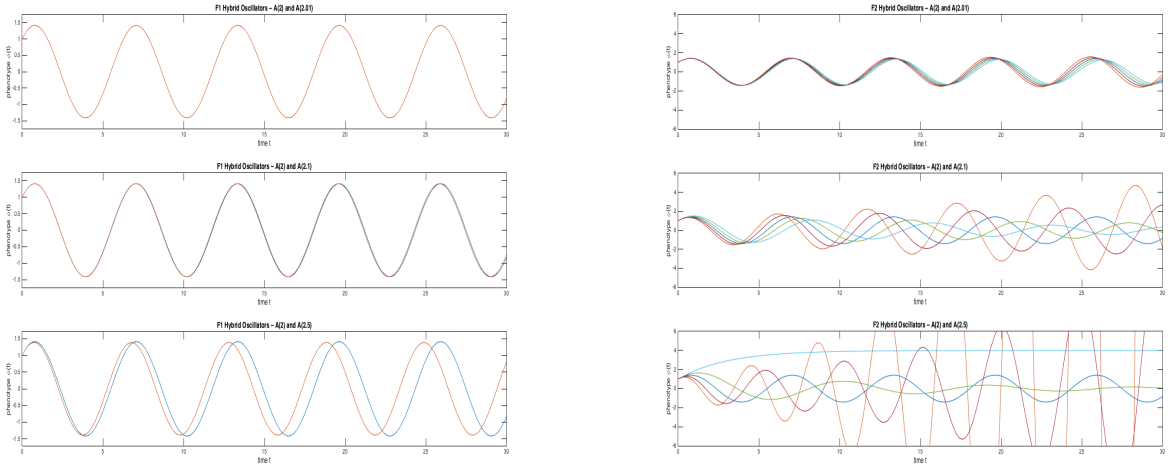

Figure 4: **(left)** Phenotypes of F_1 hybrids between an $A(2)$ parent and, top-to-bottom, an $A(2.01)$, an $A(2.1)$, and $A(2.5)$ parent. Parental phenotypes ($\sin t + \cos t$) are shown in blue, and hybrid phenotypes in orange. **(right)** Phenotypes of F_2 hybrids between the same set of parents, with parental phenotype again in blue. Different lines correspond to different F_2s produced by recombination; note that some completely fail to oscillate. *make axis labels bigger*

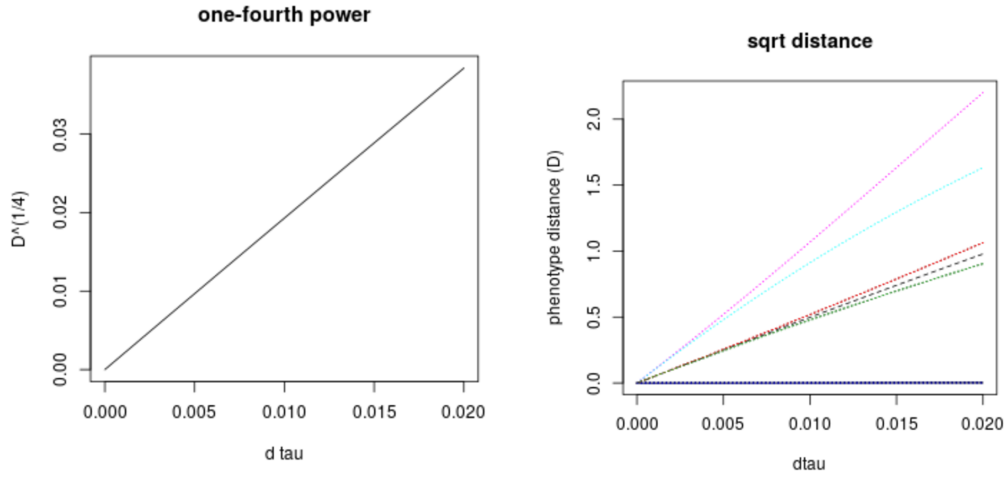

Figure 5: **(left)** Phenotype distance to optimum, D, using $\rho(t) = \exp(-t/4\pi)$ for F_1 hybrids between an $A(2)$ and an $A(2+\epsilon)$ parent. **(right)** Same, but for F_2 offspring (one line per type of F_2). *make this D not D^2 and say which line is the overall average F_2.*

System drift and the accumulation of incompatibilities

Thus far we have shown that many distinct molecular mechanisms can realize identical phenotypes and that these mechanisms may fail to produce viable hybrids. This begs the question: does evolution shift molecular mechanisms fast enough to be a significant driver of speciation? To approach this question, we explore a general quantitative genetic model in which a population drifts stochastically near a set of equivalent and optimal systems due to the action of recombination, mutation, and demographic noise. Although this is motivated by the results on linear systems above, the quantitative genetics calculations are more general, and only depend on the presence of genetic variation and a continuous set of phenotypically equivalent systems.

We will suppose that each organism's phenotype is determined by her vector of coefficients, denoted by $x = (x_1, x_2, \ldots, x_L)$, and that the corresponding fitness is determined by the distance of her phenotype to optimum. The optimum phenotype is unique, but is realized by many distinct x – those falling in the "optimal set" $\mathcal{N}$. The phenotypic distance to optimum of an organism with coefficients x is denoted $D(x)$. In the results above, $x = (A, B, C)$ and $D(x)$ is given by equation (5). determines a system (this is a system (A, B, C) above) with fitness determined by its phenotypic distance $D(x)$ from an optimum (this is analogous to $\mathcal{N}$ above). Concretely, we define the fitness at x to be $\exp(-D(x)^2)$. We will assume that in the region of interest, the map D is smooth and that we can locally approximate the optimal set $\mathcal{N}$ as a quadratic surface. As above, an individual's coefficients are given by averaging her parentally inherited coefficients and adding random noise due to segregation. Concretely, we use the *infinitesimal model* for reproduction [**?**] – the offspring of parents at x and x' will have coefficients $(x + x')/2 + \varepsilon$, where ε is a Gaussian displacement due to random assortment of parental alleles.

System drift We work with a randomly mating population of effective size N_e. If the regulatory coefficient population variation has standard deviation σ in a particular direction, since subsequent generations resample from this diversity, the population mean coefficient will move a random distance of size $\sigma/\sqrt{N_e}$ per generation, simply because this is the the standard deviation of the mean of a random sample [**?**]. Selection will tend to restrain this motion, but movement along the optimal set $\mathcal{N}$ is unconstrained, and so we expect the population mean to drift along the optimal set like a particle diffusing. The amount of variance in particular directions in

coefficient space depends on constraints imposed by selection and correlations between the genetic variation underlying different coefficients (the G matrix [**?**]). It therefore seems reasonable to coarsely model the time evolution of population variation in regulatory coefficients as a "cloud" of width σ about the population mean, which moves as an unbiased Brownian motion through the set of network coefficients that give the optimal phenotype.

There will in general be different amounts of variation in different directions; to keep the discussion intuitive, we only discuss σ_N, the amount of variation in "neutral" directions (i.e., directions along $\mathcal{N}$), and σ_S, the amount of variation in "selected" directions (perpendicular to $\mathcal{N}$). The other relevant scale we denote by γ, which the scale on which distance to phenotypic optimum changes as x moves away from the optimal set, $\mathcal{N}$. Concretely, γ is $1/(\frac{d}{du}D(x+uz))$ with respect to u where x is optimal and z is a "selected" direction perpendicular to $\mathcal{N}$. With these parameters, a typical individual will have a fitness of around $\exp(-(\sigma_S/\gamma)^2)$. Of course, there are in general many possible neutral and selected directions; we take these values to be representative of the possible directions.

Hybridization The means of two allopatric populations each of effective size N_e separated for T generations will be a distance roughly of order $2\sigma_N\sqrt{T/N_e}$ apart along $\mathcal{X}$. (Consult figure 3 for a conceptual diagram.) A population of F_1 hybrids has one haploid genome from each, whose coefficients are averaged, and so will have mean system coefficients at the midpoint between their means. Each F_2 hybrid will be homozygous for one parental allele on average at half of the loci in the genome, so the distribution of F_2s will have mean at the average of the two populations, but will have higher variance. Concretely, we expect the population of F_1s to have variance σ_S^2 in the selected direction (the same as within each parental population), but the population of F_2s will have variance of order $\sigma_S^2 + \sigma_H^2 T/N_e$, where σ_H^2 is a factor that depends on the genetic basis of the coefficients. *say something about σ_H?*

What are the fitness consequences? If a population has a Gaussian distribution of trait values with variance σ^2 and fitness decays as $\exp(-x^2/2\gamma)$ away from the population mean, then the population mean fitness is

$$\int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{x^2}{2\gamma^2}} dx = \sqrt{\frac{1}{1+(\sigma/\gamma)^2}} \leq \frac{1}{\sigma/\gamma}.$$

If σ is much smaller than γ, the fitness consequences of variation (i.e., genetic load) are small, but if σ is of order γ, then mean fitness is inversely proportional to trait variance (i.e., γ/σ). This implies that we expect $\sigma_S < \gamma$ but that once T is of order $N_e(\gamma/\sigma_H)^2$, there will be loss of fitness in F_2 hybrids. Said another way, if we write $\mathcal{F}(T)$ as the mean fitness of an F_2 hybrid between two populations separated by T generations, then fitness drops with the square root of T, measured in units of N_e generations:

$$\mathcal{F}(T) \leq \frac{1}{\sigma_H/\gamma}\sqrt{\frac{N_e}{T}}.$$

The fitness consequences of system drift above were due to increased *variance* in F_2 hybrids. If the optimal set $\mathcal{N}$ is curved (which it will be in general), then the mean F_1 offspring will fall away from the optimum slightly, leading to an additional fitness cost in hybrids. However, as shown in appendix D and depicted in figure 5, the deviation of the mean from the optimum grows proportionally to T/N_e.

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{-\frac{x^2}{2\gamma^2}} dx = \sqrt{\frac{1}{1+\sigma^2/\gamma^2}} \exp\left\{-\frac{\mu^2}{\gamma^2}\left(\frac{1}{1+\sigma^2/\gamma^2}\right)\right\}$$

With $\sigma = c_\sigma\gamma\sqrt{T/N_e}$ and $\mu = c_\mu\gamma(T/N_e)^2$, this is

$$\sqrt{\frac{1}{1+c_\sigma^2 T/N_e}} \exp\left\{-c_\mu^2\frac{T^2}{N_e^2}\left(\frac{1}{1+c_\sigma^2 T/N_e}\right)\right\}$$

improve figure by putting labels on from the following Suppose that two populations have drifted independently to differ by z, and that z is of the same order as σ but is smaller than γ. The mean F_1 is the average of the parental means, and since the first-order terms in the Taylor series vanish, has phenotype differing from the optimum by a distance of order $\|z\|^2$ (see appendix D). The mean F_2 is the same, but the standard deviation is of order z, so that up to lower order terms, while the typical fitness of an individual in the original population is $\mathcal{F}_0 = \exp(-(\sigma_s/\gamma)^2)$; of an F_1 is $\mathcal{F}_1/\mathcal{F}_0 = \exp(-(c_1\sigma_N^2 T/N_e)^2)$; and of an F_2 is $\mathcal{F}_2/\mathcal{F}_0 = \exp(-T/(N_e\gamma^2))$.

Speciation rates under neutrality Assume that perpendicular variation segregating within a population is constrained by selection such that $\sigma_S/\gamma = \xi$, and that neutral variation is on the order of perpendicular variation, such that $\sigma_N = \sigma_S$. Since the typical distance between two individuals drawn from separate allopatric populations will be $z = 2\sigma_N\sqrt{T/N_e}$, we expect a drop in hybrid fitness of

$$\mathcal{F}_2/\mathcal{F}_0 = \exp\left(-4\xi^2\frac{T}{N_e}\right). \tag{6}$$

Hybrid fitness depression will be significant when $z > \gamma$. If ξ is in $[1/5, 1]$, speciation will occur on the order of N_e generations, and possibly much faster. Note, however, that if intrapopulation variation is very small such that $\sigma_S \ll \gamma$, speciation rates will be slower.

insert commented-out discussion of what the actual genetic basis is

Discussion

The complexity of biological systems has limited our understanding of their function and evolution. Above we outline an approach, a first step, towards untangling this complexity in reference to function and evolution. This methodology borrows successfully applied tools from engineering and aims to synthesize these with the concepts and tools of molecular and evolutionary biology.

Theoretical models in evolution and population genetics often lack the molecular details of physiology or of the genotype-phenotype map. Here, we offer a tractable and simple model which includes these missing features. Further, we provide, in clear mathematical language, an analytical description of phenomena hitherto discussed verbally and conceptually (phenogenetic drift [Weiss and Fullerton, 2000], developmental systems drift [True and Haag, 2001], biological degeneracy [Edelman and Gally, 2001], *etc.*). The tractability and relative simplicity of this exposition enables the interested biologist to work out by hand, if desired, the dynamics of a genetic system, as well as perturbations to the system – an attribute not likely to be found in less tractable models and simulations.

We have suggested an interpretation of system identification: to see it as an evolutionarily neutral manifold, and not simply a computational nuisance. We have demonstrated a method to analytically determine the set of all phenotypically invariant gene networks; by a simple change of coordinates in the minimal configuration, or more generally by applying the Kalman decomposition in higher dimensions. This set is explored over evolutionary time when phenotype is conserved, and can lead to a diverse set of consequences, including the accumulation of Dobzhansky-Muller incompatibilities. *check original Bateson/DM papers to see if this accords with those defnsI read through Orr's review of those papers, which included excerpts. It seems like the DMI definition is very general and this accords.* We emphasize that these incompatibilities are a consequence of recombining different, yet functionally equivalent, mechanisms. *See refs in Barton 2010 paper.*

Furthermore, using a quantitative genetic approach, we estimated that a genetically variable population will drift in neutral system space at a rate determined by its intra-population variation and its effective population size. Because mechanistically distinct yet phenotypically equivalent biological systems can fail to produce viable hybrids, we predict allopatric populations to accumulate genetic incompatibilities at a rate on the order of N_e if intrapopulation regulatory variation is constrained by selection. Additionally we see second-generation hybrid fitness plummet much faster than that of first-generation hybrids. *refer to Turelli*

here This result is also consistent with Haldane's rule; that if only one hybrid sex is inviable or sterile it is likely the heterogametic sex. The consistency comes from gene networks localized to the sex chromosomes functioning as an F_2 hybrid cross within a diploid F_1 heterogamete as there is only one sex chromosome.

Lastly, we show that hybrid gene networks, under neutral processes, break down as function of genetic distance, and may, in part, explain broad patterns of reproductive isolation among diverse phyla [Roux et al., 2016].

Note that this follows directly from the fact that F_1 and F_2 hybrid phenotypes diverge from F_0 phenotypes, linearly and inverse-quadratically, with respect to time (more precisely T/N_e). As such, removing assumptions about the form of the fitness function, this model still predicts hybrid F_2 phenotypes to diverge much faster than F_1s, initially. So even if fitness isn't as above maybe you can see this in phenotypes.

What about nonlinearity?

Note that the drift above could actually be caused by the optimal phenotype shifting.

Acknowledgements

We would like to thank Sergey Nuzhdin, Stevan Arnold, Erik Lundgren, and Hossein Asgharian for valuable discussion.

References

Md Kausar Alam and Susan GW Kaminskyj. Aspergillus galactose metabolism is more complex than that of saccharomyces: the story of galdgal7 and galegal1. *Botany*, 91(7):467–477, 2013. 2

BDO Anderson, RW Newcomb, RE Kalman, and DC Youla. Equivalence of linear time-invariant dynamical systems. *Journal of the Franklin Institute*, 281(5):371–378, 1966. 2, 4

Richard Ernest Bellman and Karl Johan Åström. On structural identifiability. *Mathematical biosciences*, 7 (3-4):329–339, 1970. 1

Aviv Bergman and Mark L Siegal. Evolutionary capacitance as a general feature of complex gene networks. *Nature*, 424(6948):549–552, 2003. 1

Aleksandra A. Chertkova, Joshua S. Schiffman, Sergey V. Nuzhdin, Konstantin N. Kozlov, Maria G. Samsonova, and Vitaly V. Gursky. In silico evolution of the drosophila gap gene regulatory sequence under elevated mutational pressure. *BMC Evolutionary Biology*, 17(1):4, 2017. ISSN 1471-2148. doi: 10.1186/s12862-016-0866-y. URL http://dx.doi.org/10.1186/s12862-016-0866-y. 1

Ulla Christensen, Birgit S Gruben, Susan Madrid, Harm Mulder, Igor Nikolaev, and Ronald P de Vries. Unique regulatory mechanism for d-galactose utilization in aspergillus nidulans. *Applied and environmental microbiology*, 77(19):7084–7087, 2011. 2

Gheorghe Craciun and Casian Pantea. Identifiability of chemical reaction networks. *Journal of Mathematical Chemistry*, 44(1):244–259, 2008. 2

Anton Crombach, Karl R Wotton, Eva Jiménez-Guri, and Johannes Jaeger. Gap gene regulatory dynamics evolve along a genotype network. *Molecular biology and evolution*, 33(5):1293–1307, 2016. 1

Frederick R Cross, Nicolas E Buchler, and Jan M Skotheim. Evolution of networks and sequences in eukaryotic cell cycle control. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1584):3532–3544, 2011. 2

Chiraj K Dalal, Ignacio A Zuleta, Kaitlin F Mitchell, David R Andes, Hana El-Samad, and Alexander D Johnson. Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *Elife*, 5:e18981, 2016. 2

Eric H Davidson and Douglas H Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006. 1

Jeremy Draghi and Michael Whitlock. Robustness to noise in gene expression evolves despite epistatic constraints in a model of gene networks. *Evolution*, 69(9):2345–2358, 2015. 1

Gerald M Edelman and Joseph A Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24):13763–13768, 2001. 1, 11

M Grewal and K Glover. Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on automatic control*, 21(6):833–837, Dec 1976. doi: 10.1109/TAC.1976.1101375. 1

Thomas F. Hansen and Emilia P. Martins. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4):1404–1417, 1996. ISSN 00143820, 15585646. URL http://www.jstor.org/stable/2410878. 18, 23

Emily E Hare, Brant K Peterson, Venky N Iyer, Rudolf Meier, and Michael B Eisen. Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet*, 4 (6):e1000106, 2008. 2

Lukas Hartl, Christian P Kubicek, and Bernhard Seiboth. Induction of the gal pathway and cellulase genes involves no transcriptional inducer function of the galactokinase in hypocrea jecorina. *Journal of Biological Chemistry*, 282(25):18654–18659, 2007. 2

Jennifer W Israel, Megan L Martik, Maria Byrne, Elizabeth C Raff, Rudolf A Raff, David R McClay, and Gregory A Wray. Comparative developmental transcriptomics reveals rewiring of a highly conserved gene regulatory network during a major life history switch in the sea urchin genus heliocidaris. *PLoS Biol*, 14 (3):e1002391, 2016. 1

Johannes Jaeger. The gap gene network. *Cellular and Molecular Life Sciences*, 68(2):243–274, 2011. 1

Johannes Jaeger, Svetlana Surkova, Maxim Blagov, Hilde Janssens, David Kosman, Konstantin N Kozlov, Ekaterina Myasnikova, Carlos E Vanario-Alonso, Maria Samsonova, David H Sharp, et al. Dynamic control of positional information in the early drosophila embryo. *Nature*, 430(6997):368–371, 2004. 1

Johannes Jaeger, Manfred Laubichler, and Werner Callebaut. The comet cometh: evolving developmental systems. *Biological theory*, 10(1):36–49, 2015. 1

R. E. 1930-(Rudolf Emil) Kalman, Peter L. Falb, and Michael A. Arbib. *Topics in mathematical system theory*. McGraw-Hill, New York, 1969. ISBN 0754321069. 4

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *J.S.I.A.M Control*, 1963. 2, 4

Stephen E Kearsey and Sue Cotterill. Enigmatic variations: divergent modes of regulating eukaryotic dna replication. *Molecular cell*, 12(5):1067–1075, 2003. 2

Konstantin Kozlov, Svetlana Surkova, Ekaterina Myasnikova, John Reinitz, and Maria Samsonova. Modeling of gap gene expression in *Drosophila Kruppel* mutants. *PLoS Computational Biology*, 2012. 1

Konstantin Kozlov, Vitaly Gursky, Ivan Kulakovskiy, and Maria Samsonova. Sequnce-based model of gap gene regulatory network. *BMC Genomics*, 2014. 1

Konstantin Kozlov, Vitaly V Gursky, Ivan V Kulakovskiy, Arina Dymova, and Maria Samsonova. Analysis of functional importance of binding sites in the *Drosophila* gap gene network model. *BMC Genomics*, 2015. 1

Russell Lande. Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences*, 78(6):3721–3725, 1981. URL http://www.pnas.org/content/78/6/3721.abstract. 18

Hugo Lavoie, Hervé Hogues, and Malcolm Whiteway. Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi. *Current opinion in microbiology*, 12(6):655–663, 2009. 2

Mikhail Martchenko, Anastasia Levitin, Herve Hogues, Andre Nantel, and Malcolm Whiteway. Transcriptional rewiring of fungal galactose-metabolism circuitry. *Current Biology*, 17(12):1007–1013, 2007. 2

Eric Mjolsness, David H Sharp, and John Reinitz. A connectionist model of development. *Journal of theoretical Biology*, 152(4):429–453, 1991. 1

Mihaela Pavlicev and Gunter P Wagner. A model of developmental evolution: selection, peliotropy and compensation. *Trends in Ecology & Evolution*, 2012. 1

Camille Roux, Christelle Fraisse, Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, and Nicolas Bierne. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS biology*, 14 (12):e2000234, 2016. 12

Aziz Sancar. The intelligent clock and the rube goldberg clock. *Nature structural & molecular biology*, 15 (1):23–24, 2008. 2

Maria R Servedio, Yaniv Brandvain, Sumit Dhole, Courtney L Fitzpatrick, Emma E Goldberg, Caitlin A Stern, Jeremy Van Cleve, and D Justin Yeh. Not just a theorythe utility of mathematical models in evolutionary biology. *PLoS Biol*, 12(12):e1002017, 2014. 1

Mark L Siegal and Aviv Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences*, 99(16):10528–10532, 2002. 1

Andrew B. Stergachis, Shane Neph, Richard Sandstrom, Eric Haugen, Alex P. Reynolds, Miaohua Zhang, Rachel Byron, Theresa Canfield, Sandra Stelhing-Sun, Kristen Lee, Robert E. Thurman, Shinny Vong, Daniel Bates, Fidencio Neri, Morgan Diegel, Erika Giste, Douglas Dunn, Jeff Vierstra, R. Scott Hansen, Audra K. Johnson, Peter J. Sabo, Matthew S. Wilken, Thomas A. Reh, Piper M. Treuting, Rajinder Kaul, Mark Groudine, M. A. Bender, Elhanan Borenstein, and John A. Stamatoyannopoulos. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–370, November 2014. ISSN 00280836. URL http://dx.doi.org/10.1038/nature13972. 2

John R True and Eric S Haag. Developmental system drift and flexibility in evolutionary trajectories. *Evolution & development*, 3(2):109–119, 2001. 1, 2, 11

Annie E Tsong, Brian B Tuch, Hao Li, and Alexander D Johnson. Evolution of alternative transcriptional circuits with identical logic. *Nature*, 443(7110):415–420, 2006. 2

AJ Van der Schaft. Equivalence of dynamical systems by bisimulation. *IEEE transactions on automatic control*, 49(12):2160–2172, 2004. 2

Andreas Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences*, 91(10):4387–4391, 1994. 1

Andreas Wagner. Does evolutionary plasticity evolve? *Evolution*, pages 1008–1023, 1996. 1

Eric Walter, Yves Lecourtier, and John Happel. On the structural output distinguishability of parametric models, and its relations with structural identifiability. *IEEE Transactions on Automatic Control*, 29(1): 56–57, 1984. 1

Kenneth M Weiss and Stephanie M Fullerton. Phenogenetic drift and the evolution of genotype–phenotype relationships. *Theoretical population biology*, 57(3):187–195, 2000. 1, 11

Karl R Wotton, Eva Jiménez-Guri, Anton Crombach, Hilde Janssens, Anna Alcaine-Colet, Steffen Lemke, Urs Schmidt-Ott, and Johannes Jaeger. Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly megaselia abdita. *Elife*, 4:e04785, 2015. 1

Lotfi A Zadeh and Charles A Deoser. *Linear system theory*. Robert E. Krieger Publishing Company Huntington, 1976. 2

Examples

Example 4 (Metabolic network)**.** *Consider an organism that can metabolize two different sugars (present at logarithmic concentrations u_1 and u_2), with two enzymes (at log concentrations ϕ_1 and ϕ_2). Further suppose that the second sugar is prefered, and that depending on u_1 and u_2 the organism will synthesize an appropriate ϕ_1 and ϕ_2. Furthermore, assume this system contains at least two transcription factors, whose log concentrations are κ_1, κ_2, $\dots \kappa_n$. Minimally such a system may have the architecture, $\mathcal{S}_{\min}(U) = (U \left[\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right] U^{-1}, U, U^{-1})$. We can find alternative equivalent systems by changning coordinates ($U \to U'$) or, more generally, by applying the Kalman decomposition A. To illustrate that phenotypic invariance does not require dimensional invariance, we apply the Kalman decomposition to $\mathcal{S}_{\min}$ to find $\mathcal{S}(D_{1-3}, V) = (\, V \left[\begin{smallmatrix} D_1 & D_2 \\ 0 & A \end{smallmatrix}\right] V^{-1}, V \left[\begin{smallmatrix} D_3 \\ B \end{smallmatrix}\right], [0 \; C] V^{-1} \,)$, both of which are in $\mathcal{N}$ (V can be any 4-dimensional and invertible matrix, $D_{1-3} \in \mathbb{R}^{2\times 2}$, and $(A, B, C) \in \mathcal{S}_{\min}$).*

A Kalman Decomposition

Definition 1 (Phenotypic equivalence of systems)**.** *Let $(\kappa(t), \phi(t))$ and $(\bar{\kappa}(t), \bar{\phi}(t))$ be the solutions to (??) with coefficient matrices (A, B, C) and $(\bar{A}, \bar{B}, \bar{C})$ respectively, and both $\kappa(0)$ and $\bar{\kappa}(0)$ are zero. The systems defined by (A, B, C) and $(\bar{A}, \bar{B}, \bar{C})$ are* ***phenotypically equivalent*** *if*

$$\phi(t) = \bar{\phi}(t) \qquad \text{for all } t \geq 0.$$

Equivalently, this occurs if and only if

$$h(t) = \bar{h}(t) \qquad \text{for all } t \geq 0,$$

where h and $\bar{h}$ are the impulse responses of the two systems.

One way to find other systems equivalent to a given one is by change of coordinates ("algebraic equivalence"): if V is an invertible matrix, then the systems (A, B, C) and (VAV^{-1}, VB, CV^{-1}) have the same dynamics because their transfer functions are equal:

$$CV^{-1}(zI - VAV^{-1})^{-1}VB = CV^{-1}V(zI - A)^{-1}V^{-1}VB = C(zI - A)^{-1}B.$$

However, the converse is not necessarily true: systems can have identical transfer functions without being changes of coordinates of each other. In fact, systems with identical transfer functions can involve interactions between different numbers of molecular species.

The set of all systems phenotypically equivalent to a given system (A, B, C) is elegantly described using the Kalman decomposition, which also clarifies the system dynamics? tells us a lot about how it works? *or something* To motivate this, first note that the input $u(t)$ only directly pushes the system in directions lying in the span of the columns of B. As a result, different combinations of input can move the system in any direction that lies in the *reachable subspace*, which we denote by $\mathcal{R}$, and is defined to be the closure of span(B) under applying A (or equivalently, the span of $B, AB, A^2B, \dots A^{n-1}B$). Analogously to this, we

define the *observable subspace*, $\mathcal{O}$, to be the closure of span(C^T) under applying A. (Or: $\bar{\mathcal{O}}$ is the largest A-invariant subspace contained in the null space of C; and $\mathcal{R}$ is the largest A-invariant subspace contained in the image of B.)

If we define

1. The columns of $P_{r\bar{o}}$ are an orthonormal basis for $\mathcal{R} \cap \bar{\mathcal{O}}$.
2. The columns of P_{ro} are an orthonormal basis of the complement of $\mathcal{R} \cap \bar{\mathcal{O}}$ in $\mathcal{R}$.
3. The columns of $P_{\bar{r}o}$ are an orthonormal basis of the complement of $\mathcal{R} \cap \bar{\mathcal{O}}$ in $\bar{\mathcal{O}}$.
4. The columns of $P_{\bar{r}\bar{o}}$ are an orthonormal basis of the remainder of $\mathbb{R}^n$.

If we then define

$$P = \left[\begin{array}{c|c|c|c} P_{r\bar{o}} & P_{ro} & P_{\bar{r}o} & P_{\bar{r}\bar{o}} \end{array} \right],$$

then

$$P^T P = \left[\begin{array}{c|c|c|c} I & 0 & 0 & 0 \\ \hline 0 & I & U & 0 \\ \hline 0 & V & I & 0 \\ \hline 0 & 0 & 0 & I \end{array}\right].$$

Check this. Can we get $U = V = 0$?

The following theorem can be found in SOME REFERENCE.

Theorem 1 (Kalman decomposition)**.** *For any system (A, B, C) with corresponding Kalman basis matrix P, the transformed system (PAP^{-1}, PB, CP^{-1}) has the following form:*

$$\widehat{A} = PAP^{-1} = \left[\begin{array}{cccc} A_{r\bar{o}} & A_{r\bar{o},ro} & A_{r\bar{o},\bar{r}\bar{o}} & A_{r\bar{o},\bar{r}o} \\ 0 & A_{ro} & 0 & A_{ro,\bar{r}o} \\ 0 & 0 & A_{\bar{r}\bar{o}} & A_{\bar{r}\bar{o},\bar{r}o} \\ 0 & 0 & 0 & A_{\bar{r}o} \end{array}\right],$$

and

$$\widehat{B} = PB = \left[\begin{array}{c} B_{r\bar{o}} \\ B_{ro} \\ 0 \\ 0 \end{array}\right],$$

and

$$\widehat{C} = CP^{-1} = \left[\begin{array}{cccc} 0 & C_{ro} & C_{\bar{r}\bar{o}} & 0 \end{array}\right].$$

The transfer function of both systems is given by

$$H(z) = C_{ro}(zI - A_{ro})^{-1} B_{ro}.$$

In the latter case, we say that the system is *minimal* – there is no equivalent system with a smaller number of species. Note that this says that any two equivalent minimal systems are changes of basis of each other.

Since any system can be put into this form, and once in this form, its transfer function is determined only by C_{ro}, A_{ro}, and B_{ro}, therefore, the set of all equivalent systems are parameterized by the dimension n, the choice of basis (P), the remaining submatrices in $\widehat{A}$, $\widehat{B}$, and $\widehat{C}$ (which are unconstrained), and a invertible transformation of span(P_{ro}), which we call T_{ro}.

Theorem 2 (Parameterization of equivalent systems)**.** *Let (A, B, C) be a minimal system.*

(a) Every equivalent system is of the form given in Theorem 1, i.e., can be specified by choosing a dimension, n; submatrices in $\widehat{A}$, $\widehat{B}$, and $\widehat{C}$ except for $A_{ro} = A$, $B_{ro} = B$, and $C_{ro} = C$; and choosing an invertible matrix P.

(b) conjecture: The parameterization is unique if P is furthermore chosen so that each P_x other than P_{ro} is a projection matrix, and that

$$0 = P_x^T P_y$$

for all (x, y) except $(ro, \bar{r}\bar{o})$.

Another way of saying it: pick the $\mathcal{R}$ and $\bar{\mathcal{O}}$ subspaces, that must intersect in something of the minimal dimension; then let P be the appropriate basis?

In some situations we may be interested in only "network rewiring", where A changes while B and C do not. For instance, if all non-regulatory functions of each molecule are strongly constrained, then C cannot change. Likewise, if responses of each molecule to the external inputs are not changed by evolution, then B does not change.

A.1 Neutral directions from the Kalman decomposition

The Kalman decomposition above says that any system (A, B, C) can be decomposed into $(P, \hat{A}, \hat{B}, \hat{C})$ so that

$$\begin{aligned} A &= P^{-1}\hat{A}P \\ B &= P^{-1}\hat{B} \\ C &= \hat{C}P, \end{aligned}$$

and we know precisely how we can change these to preserve the transfer function:

1. $P \to P + \epsilon Q$ as long as the result is still invertible,
2. $\hat{A} \to A + \epsilon X$ as long as X is zero in the correct places,
3. $\hat{B} \to B + \epsilon Y$ as long as Y is zero in the correct places,
4. $\hat{C} \to C + \epsilon Y$ as long as Z is zero in the correct places.

By taking $\epsilon \to 0$, these tell us the local directions we can move a system (A, B, C) in. All statements below are up to first order in ϵ, omitting terms of order ϵ^2.

First, since $(P + \epsilon Q)^{-1} = P^{-1} + \epsilon P^{-1} Q P^{-1}$, modifying $P \to P + \epsilon Q$ changes

$$\begin{aligned} A &\to A + \epsilon P^{-1}\hat{A}Q - \epsilon P^{-1}QP^{-1}\hat{A}P \\ &= A + \epsilon\left(AP^{-1}Q - P^{-1}QA\right), \\ B &\to B - \epsilon P^{-1}QB \\ C &\to C + \epsilon CP^{-1}Q. \end{aligned}$$

Since P is invertible and Q can be anything (if ϵ is small enough), this allows changes in the direction of an arbitrary W:

$$\begin{aligned} A &= A + \epsilon\left(AW - WA\right), \\ B &\to B - \epsilon WB \\ C &\to C + \epsilon CW. \end{aligned}$$

Then, $\hat{A} \to A + \epsilon X$ does

$$A \to A + \epsilon P^{-1} X P$$

and $\hat{B} \to B + \epsilon Y$ does

$$B \to B + \epsilon P^{-1} Y$$

and $\hat{C} \to C + \epsilon Z$ does

$$C \to C + \epsilon Z P.$$

These degrees of freedom look like they depend on P, which is not unique, but for any two choices of P there are corresponding choices of X that give the same actual change in A (and likewise for Y and Z).

Therefore, this gives us an upper bound on the number of degrees of freedom, in terms of the dimensions of the blocks in the Kalman decomposition (n_{ro} etc) and the dimensions of B and C (n_B and n_C respectively): namely, for W, X, Y, and Z respectively:

$$n^2 + (n_{r\bar{o}} + n_{ro} n_{\bar{r}o} + n_{\bar{r}\bar{o}}(n_{\bar{r}\bar{o}} + n_{\bar{r}o}) + n_{\bar{r}o}^2) + n_B n_{r\bar{o}} + n_C n_{\bar{r}\bar{o}}.$$

However, some of these may be redundant. For instance, changing P in the direction of a Q that satisfies both $AP^{-1}Q = P^{-1}QA$ and $CP^{-1}Q = 0$ is equivalent to changing B by $Y = QB$.

B Meiotic recombination in linear systems

Recombination is performed by taking two analogous system components from $\mathcal{S}$ and $\mathcal{S}'$ and randomly swapping rows or columns. *E.g.* gamete systems (A'', B'', C''), produced by the diploid $\{\mathcal{S}, \mathcal{S}'\}$, are formed by recombining (randomly swapping rows or columns) between two, possibly distinct, systems $\mathcal{S} = (A, B, C)$ and $\mathcal{S}' = (A', B', C')$ such that,

$$\mathcal{S}'' = \begin{pmatrix} A'' & = MA + (I - M)A', \\ B'' & = MB + (I - M)B', \\ C'' & = CM + C'(I - M) \end{pmatrix}$$

where M is a diagonal matrix where each diagonal element is a Bernoulli random variable ($M_{ii} = 0$ or 1 with equal probability, and $M_{ij} = 0$ if $i \neq j$). If systems are different dimensions the smaller system elements can be augmented with 0s (*e.g.* $\left[\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} B \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} C & 0 \end{smallmatrix}\right]$).

C Genetic drift with a multivariate trait

For completeness, we provide a brief exposition of how a population evolves due to genetic drift with a quantitative genetics model, as in Lande [1981] or Hansen and Martins [1996]. These do not directly model underlying genetic basis, but developing a more accurate model is beyond the scope of this paper.

Suppose that the population is distributed in trait space as a Gaussian with covariance matrix Σ and mean μ, whose density we write as $f(\cdot; \Sigma, \mu)$. Selection has the effect of multiplying this density by the fitness function and renormalizing, so that if expected fitness of x is proportional to $\exp(-\|Lx\|^2/2)$, then the distribution post-selection has density at x proportional to $f(x; \Sigma, \mu)\exp(-\|Lx\|^2/2)$. By the computation below ("Completing the square"), the result is a Gaussian distribution with covariance matrix $(\Sigma^{-1} + L^T L)^{-1}$ and mean $(\Sigma^{-1} + L^T L)^{-1}\Sigma^{-1}\mu$.

After selection, we have reproduction: suppose this occurs as in the infinitesimal model [**?**], so that each offspring of parents with traits x and y is drawn independently from a Gaussian distribution with mean $(x + y)/2$ and covariance matrix R. Here, R is the contribution of "segregation variance". If $\widetilde{\Sigma} = (\Sigma^{-1} + L^T L)^{-1}$ is the covariance matrix of the parents post-selection, then the distribution of offspring will again be Gaussian, with mean equal to that of the parents and covariance matrix $\widetilde{\Sigma}/2 + R$.

In summary, a generation under this model modifies the mean (μ) and covariance matrix (Σ) of a population as follows:

$$\mu \mapsto \mu' = (\Sigma^{-1} + L^T L)^{-1}\Sigma^{-1}\mu$$
$$\Sigma \mapsto \Sigma' = \frac{1}{2}(\Sigma^{-1} + L^T L)^{-1} + R.$$

What measures are stable under this transformation? The condition $\mu = \mu'$ reduces to $\Sigma L^T L\mu = 0$; if we assume R and therefore Σ are of full rank, then this happens if and only if μ is in the null space of L, i.e., if μ lies in a neutral direction. The condition $\Sigma' = \Sigma$ can also be solved, at least numerically. After rearrangement, it reduces to $\Sigma L^T L\Sigma + (I/2 - RL^T L)\Sigma = R$. We can find a more explicit description if we assume that $x^T L^T Lx = \sum_{i=1}^k x_i^2$, i.e., that selection only cares about the first k coordinates, and then with no interactions between traits. If so, the condition $\Sigma' = \Sigma$ can be written in block form as

$$\begin{bmatrix} \Sigma_{11}^2 + (I/2 - R_{11})\Sigma_{11} & \Sigma_{11}\Sigma_{12} + (I/2 - R_{11})\Sigma_{12} \\ \Sigma_{12}^T\Sigma_{11} + \Sigma_{12}^T/2 - R_{12}^T\Sigma_{11} & \Sigma_{22} - R_{12}^T\Sigma_{12} \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix}.$$

The first equation, $\Sigma_{11}^2 + (I/2 - R_{11})\Sigma_{11} = R_{11}$, can be solved with the quadratic formula:

$$\Sigma_{11} = (R_{11} - I/2 + Q)/2$$

for any Q that commutes with R_{11} and is a solution to $Q^2 = (R_{11} - I/2)^2 + 4R_{11}$. Since we need Σ to be positive definite, we take the solution with positive eigenvalues. Given Σ_{11}, the remaining components are

$$\Sigma_{12} = (\Sigma_{11} + I/2 - R_{11})^{-1} R_{12}$$
$$\Sigma_{22} = R_{12}^T\Sigma_{12} + R_{22}.$$

Importantly, the mean μ does not affect either how the covariance matrix moves, or its stable shape.

Above we have described the *expected* motion of the mean and covariance.. However, random resampling will introduce noise. Suppose that a population of N individuals behaves approximately as described above. By the above, we may expect that the covariance matrix stays close to a constant value Σ, computed from R and L as above, so that we need only consider motion of the mean, μ. Since we take a sample of size N to construct the next generation, the next generation's mean is drawn from a Gaussian distribution with mean μ' and covariance matrix Σ/N. Defining $\Gamma = (I - (I + \Sigma L^T L)^{-1})$, this can be written as

$$\mu' - \mu = \Gamma\mu + \epsilon/\sqrt{N},$$

where ϵ is a multivariate Gaussian with mean zero and covariance matrix Σ. Let $\mu(k)$ denote the mean in the k^{th} generation, and suppose that μ differs from optimal by something of order $1/N$: if $\nu(t) = N\mu(tN)$, then the previous equation implies that as $N \to \infty$, in the limit ν solves the Itô equation

$$d\nu(t) = \Gamma\nu(t)dt + \Sigma^{1/2}dW(t),$$

where now $W(t)$ is a multivariate white noise. This has an explicit solution as a multivariate Ornstein-Uhlenbeck process:

$$\nu(t) = e^{-t\Gamma}\nu_0 + \int_0^t e^{-(t-s)\Gamma}\Sigma^{1/2}dW(s).$$

The asymptotic variance of this process in the direction z is

$$\lim_{t\to\infty} \text{Var}[\nu(t)\cdot z] = \int_0^\infty z^T e^{-s\Gamma}\Sigma e^{-s\Gamma}z ds, \tag{7}$$

which is infinite iff $\Gamma z = 0$, which occurs iff $Lz = 0$. In other words, population mean trait values lie away from the optimal set by a Gaussian displacement of order $1/N$ with a covariance matrix given by equation (7).

Now write what this means intuitively.

Completing the square First note that if A is symmetric,

$$(x-y)^T A(x-y) = x^T A\,(x-2y) + y^T Ay,$$

and so if B is also symmetric and $A+B$ is invertible,

$$\begin{aligned}(x-y)^T A(x-y) + x^T Bx &= x^T(A+B)\left(x - 2(A+B)^{-1}Ay\right) + y^T Ay \\ &= \left(x-(A+B)^{-1}Ay\right)^T (A+B)\left(x-(A+B)^{-1}Ay\right) \\ &\quad + y^T Ay - y^T A^T (A+B)^{-1}Ay.\end{aligned}$$

¡¡¡¡¡¡¡ HEAD Since the second line doesn't depend on x, by substituting $A=\Sigma^{-1}$ and $B=U^{-1}$, ======= Therefore, by substituting $A=\Sigma^{-1}$ and $B=L^TL$, ¿¿¿¿¿¿¿ origin/master

$$\frac{f(x;\Sigma,y)\exp(-x^TL^TLx/2)}{\int f(z;\Sigma,y)\exp(-z^TL^TLz/2)dz} = f(x;(\Sigma^{-1}+L^TL)^{-1},(\Sigma^{-1}+L^TL)^{-1}\Sigma^{-1}y).$$

Evolution of segregation covariance

The description above does not completely describe how two diverging populations interact, since the amount of *segregation variance*, quantified by R, will not stay constant. To get an idea of how this might change, suppose that a multivariate trait is determined by L unlinked, biallelic loci, and that the i^{th} locus has two alleles with additive effects $\pm x_i$, so that begin homozygous for the + allele contributes $+2x_i$ to the trait. For simplicity, we will neglect the effects of selection. *fixup the below to be actually multivariate* If the + allele at locus i is at frequency p_i in a population, then the mean and genetic variance of the trait in a diploid population with random mating is *the mean should be* $\sum_i x_i(2p_i-1)$

$$\begin{aligned}m &= 2\sum_i p_i x_i \\ s^2 &= 2\sum_i p_i(1-p_i)x_i^2.\end{aligned}$$

Segregation variance between two parents depends on the loci at which either are heterozygous, and each locus contributes independently since alleles are additive. If the alleles are at Hardy-Weinberg proportions, then since segregation is a fair coin flip, a heterozygous locus contributes $x_i^2/4$ to the variance, and so the *mean* segregation variance, averaging across parents, is

$$R_0(p) = \frac{1}{2}\sum_i p_i(1-p_i)x_i^2.$$

On the other hand, if the second parent came from a distinct population with frequencies q_i (an F_1 hybrid), this would be

$$\begin{aligned}R_1(p,q) &= \frac{1}{4}\sum_i p_i^2(1-p_i)^2x_i^2 + \frac{1}{4}\sum_i q_i^2(1-q_i)^2x_i^2 \\ &= (R_0(p)+R_0(q))/2.\end{aligned}$$

If we assume that the populations are at equilibrium, $R_0(p)\approx R_0(q)$, and so $R_1(p,q)\approx R_0(p)$.

Now consider an F_2 hybrid, where both parents are F_1 and so each heterozygous at locus i with probability $p_i(1-q_i)+(1-p_i)q_i$. Then

$$R_2(p,q) = \frac{1}{4}\sum_i \left(p_i(1-q_i)+(1-p_i)q_i\right)x_i^2.$$

Suppose that the two populations are slightly drifted from each other, with frequency difference $p_i - q_i = 2\epsilon_i$. Then,

$$\begin{aligned} p(1-q) + p(1-q) &= (u+\epsilon)(1-u+\epsilon) + (u-\epsilon)(1-u-\epsilon) \\ &= 2u(1-u) + 2\epsilon^2. \end{aligned}$$

If the frequencies have evolved neutrally in unconnected, Wright-Fisher populations of effective size N for t generations from a common ancestor with allele frequency u, then ϵ has mean zero and variance roughly $u(1-u)t/N$. Still assuming the populations are at stationarity, so that R_0 is constant between the two, and taking the frequencies p_i as a proxy for the ancestral frequencies u_i, this implies that we expect

$$\begin{aligned} R_2 &\approx R_0 + \frac{1}{2}\sum_i p_i(1-p_i)x_i^2 t/N \\ &= \left(1 + \frac{t}{N}\right) R_0. \end{aligned}$$

On the other hand, the expected squared difference in trait *means* here is

$$4\sum_i p_i(1-p_i)x_i^2 t/N = 8R_0 t/N. \tag{8}$$

This implies that under this model, segregation variance in F_2s between two populations is roughly increased by a factor of 1/8 of the difference between their means.

D Away from the optimum

Let two points on $\mathcal{N}$ be x_1 and x_2, let $\bar{x} = (x_1 + x_2)/2$, and let $z = (x_2 - x_1)/2$. Then with ∇D and $\nabla^2 D$ the first and second derivatives of D, respectively, Taylor expanding about x_1 and x_2 finds that

$$\begin{aligned} D(\bar{x}) &= D(x_1) + \nabla D(x_1)\cdot z + \frac{1}{2} z^T \nabla^2 D(x_1) z + O(\|z\|^3) \\ &= D(x_2) - \nabla D(x_2)\cdot z + \frac{1}{2} z^T \nabla^2 D(x_2) z + O(\|z\|^3). \end{aligned}$$

Now, since $D(x_1) = D(x_2) = \nabla D(x_1) = \nabla D(x_2) = 0$ and

$$\begin{aligned} \nabla D(x_2) = \nabla D(x_1) + 2z^T \nabla^2 D(x_1) + O(\|z\|^2), \quad \text{and} \\ \nabla^2 D(x_2) = \nabla^2 D(x_1) + O(\|z\|), \end{aligned}$$

adding together the two equations above and dividing by two gets that

$$D(\bar{x}) = \Phi_0 - \frac{3}{2} z^T \nabla^2 D(x_1) z + O(\|z\|^3).$$

E Gaussian load

Suppose that a population has a Gaussian distribution in trait space with mean μ and covariance matrix Σ, and that fitness of an individual at x is $\exp(-\|Lx\|^2/2)$. Then, completing the square as above with

$A = \Sigma^{-1}$, $y = \mu$, and $B = L^T L$, and defining $Q = (\Sigma^{-1} + L^T L)^{-1}$,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}^n \det(\Sigma)^{1/2}} \int e^{-\frac{1}{2}x^T\Sigma^{-1}x} e^{-\frac{1}{2}x^T L^T L x} dx \\
&\quad = \frac{1}{\sqrt{2\pi}^n \det(\Sigma)^{1/2}} \int e^{-\frac{1}{2}(x-Q\Sigma^{-1}\mu)^T Q^{-1}(x-Q\Sigma^{-1}\mu)} dx \\
&\qquad \times e^{\mu^T(I-\Sigma^{-1}Q)\Sigma^{-1}\mu} \\
&\quad = \sqrt{\frac{\det(Q)}{\det(\Sigma)}} \exp\left\{\mu^T\left(I - \Sigma^{-1}Q\right)\Sigma^{-1}\mu\right\} \\
&\quad = \sqrt{\frac{1}{\det(\Sigma)\det(\Sigma^{-1} + L^T L)}} \exp\left\{\mu^T\left(I - (I + L^T L\Sigma)^{-1}\right)\Sigma^{-1}\mu\right\}
\end{aligned}$$

F Differentiating the fitness function

Add refs to sensitivity analysis.

Suppose that $\rho(t) \geq 0$ is a weighting function on $[0, \infty)$ so that fitness is a function of $L^2(\rho)$ distance of the impulse response from optimal. With $h_0(t) = C_0 e^{tA_0} B_0$ a representative of the optimal set:

$$\begin{aligned}
D(A,B,C) &:= \int_0^\infty \rho(t) \left|h_A(t) - h_0(t)\right|^2 dt \\
&:= \int_0^\infty \rho(t) \left|Ce^{At}B - C_0 e^{A_0 t} B_0\right|^2 dt \\
&= \int_0^\infty \rho(t) \operatorname{tr}\left\{\left(Ce^{At}B - C_0 e^{A_0 t} B_0\right)^T \left(Ce^{At}B - C_0 e^{A_0 t} B_0\right)\right\} dt \\
&= \int_0^\infty \rho(t) \operatorname{tr}\left\{\left(Ce^{At}B - C_0 e^{A_0 t} B_0\right) \left(Ce^{At}B - C_0 e^{A_0 t} B_0\right)^T\right\} dt.
\end{aligned} \tag{9}$$

How does this change as we perturb about (A_0, B_0, C_0)? First we differentiate with respect to A, keeping $B = B_0$ and $C = C_0$ fixed. Since

$$\frac{d}{du} e^{(A+uZ)t}|_{u=0} = \int_0^t e^{As} Z e^{A(t-s)} ds, \tag{10}$$

we have that

$$\begin{aligned}
\frac{d}{du} D(A + uZ, B_0, C_0)|_{u=0} &= 2\int_0^\infty \rho(t) \operatorname{tr}\left\{C_0 \left(\int_0^t e^{As} Z e^{A(t-s)} ds\right) B_0 B_0^T \left(e^{At} - e^{A_0 t}\right)^T C_0^T\right\} dt \\
&= 2\int_0^\infty \rho(t) \operatorname{tr}\left\{C_0 \left(\int_0^t e^{As} Z e^{A(t-s)} ds\right) B_0 \left(h_A(t) - h_0(t)\right)^T\right\} dt
\end{aligned} \tag{11}$$

and, by differentiating this and supposing that A is on the optimal set, i.e., $h_A(t) = h_0(t)$, (so wolog $A = A_0$):

$$\begin{aligned}
\mathcal{H}^{A,A}(Y,Z) &:= \frac{1}{2}\frac{d}{du}\frac{d}{dv} D(A_0 + uY + vZ, B_0, C_0)|_{u=v=0} \\
&= \int_0^\infty \rho(t) \operatorname{tr}\left\{C_0 \left(\int_0^t e^{A_0 s} Y e^{A_0(t-s)} ds\right) B_0 B_0^T \left(\int_0^t e^{A_0 s} Z e^{A_0(t-s)} ds\right)^T C_0^T\right\} dt.
\end{aligned} \tag{12}$$

The function $\mathcal{H}$ will define a quadratic form. To illustrate the use of this, suppose that B and C are fixed. By defining Δ_{ij} to be the matrix with a 1 in the (i,j)th slot and 0 elsewhere, the coefficients of the quadratic form is

$$H_{ij,k\ell}(A) := \mathcal{H}(\Delta_{ij}, \Delta_{k\ell}). \tag{13}$$

We could use this to compute the gradient of D, or to get the quadratic approximation to D near the optimal set. To do so, it'd be nice to have a way to compute the inner integral above. Suppose that we can diagonalize $A = U\Lambda U^{-1}$. Then

$$\int_0^t e^{As} Z e^{A(t-s)} ds = \int_0^t U e^{\Lambda s} U^{-1} Z U e^{\Lambda(t-s)} U^{-1} ds \tag{14}$$

Now, notice that

$$\int_0^t e^{s\lambda_i} e^{(t-s)\lambda_j} ds = \frac{e^{t\lambda_i} - e^{t\lambda_j}}{\lambda_i - \lambda_j}. \tag{15}$$

Therefore, defining

$$X_{ij}(t, Z) = \left(U^{-1} Z U\right)_{ij} \frac{e^{t\lambda_i} - e^{t\lambda_j}}{\lambda_i - \lambda_j} \tag{16}$$

moving the U and U^{-1} outside the integral and integrating we get that

$$\int_0^t e^{As} Z e^{A(t-s)} ds = U X(t, Z) U^{-1}. \tag{17}$$

Following on from above, we see that if $Z = \Delta_{k\ell}$, then

$$X_{ij}^{k\ell}(t) = \frac{e^{t\lambda_i} - e^{t\lambda_j}}{\lambda_i - \lambda_j} (U^{-1})_{\cdot k} U_{\ell \cdot}, \tag{18}$$

where $U_{k\cdot}$ is the kth row of U, and so

$$H_{ij,k\ell}(A) = \int_0^\infty \rho(t) \operatorname{tr}\left\{ C U X^{ij}(t) U^{-1} B B^T (U^{-1})^T X^{k\ell}(t)^T U^T C^T \right\} dt. \tag{19}$$

This implies that

$$D(A_0 + \epsilon Z) \approx \epsilon^2 \sum_{ijk\ell} H^{ij,k\ell} Z_{ij} Z_{k\ell} \tag{20}$$

and so

$$D(A_0 + \epsilon Z) \approx \epsilon^2 \sum_{ijk\ell} H^{ij,k\ell} Z_{ij} Z_{k\ell} \tag{21}$$

By section C, if we set $\Sigma = \sigma^2 I$ and $U = H$, then a population at $A_0 + Z$ experiences a restoring force of strength $(I + \sigma^2 H^{-1})^{-1} Z$ (treating Z as a vector and H as an operator on these). If σ^2 is small compared to H^{-1} then this is approximately $-\sigma^2 H^{-1} Z$. This suggests that the population mean follows an Ornstein-Uhlenbeck process, as described (in different terms) in Hansen and Martins [1996].

More generally, B and C may also change. To extend this we need the remaining second derivatives of D. First, in B:

$$\begin{aligned}
\mathcal{H}^{B,B}(Y, Z) &:= \frac{1}{2} \frac{d}{du} \frac{d}{dv} D(A_0, B_0 + uY + vZ, C_0)|_{u=v=0} \\
&= \frac{1}{2} \int_0^\infty \rho(t) \operatorname{tr}\left\{ C_0 e^{tA_0} \frac{d}{du} \frac{d}{dv} (uY + vZ)(uY + vZ)^T |_{u=v=0} e^{tA_0^T} C_0^T \right\} dt \\
&= \frac{1}{2} \int_0^\infty \rho(t) \operatorname{tr}\left\{ C_0 e^{tA_0} \left(YZ^T + ZY^T\right) e^{tA_0^T} C_0^T \right\} dt.
\end{aligned} \tag{22}$$

Next, in C:

$$\begin{aligned}\mathcal{H}^{B,B}(Y,Z) &:= \frac{1}{2}\frac{d}{du}\frac{d}{dv}D(A_0,B_0,C_0+uY+vZ)|_{u=v=0}\\ &= \frac{1}{2}\int_0^\infty \rho(t)\,\mathrm{tr}\left\{B_0 e^{tA_0^T}\frac{d}{du}\frac{d}{dv}(uY+vZ)^T(uY+vZ)|_{u=v=0}e^{tA_0}B_0\right\}dt\\ &= \frac{1}{2}\int_0^\infty \rho(t)\,\mathrm{tr}\left\{B_0 e^{tA_0^T}\left(YZ^T+ZY^T\right)e^{tA_0}B_0\right\}dt.\end{aligned} \tag{23}$$

Now, the mixed derivatives in B and C:

$$\begin{aligned}\mathcal{H}^{B,C}(Y,Z) &:= \frac{1}{2}\frac{d}{du}\frac{d}{dv}D(A_0,B_0+uY,C_0+vZ)|_{u=v=0}\\ &= \int_0^\infty \rho(t)\,\mathrm{tr}\left\{Ye^{tA_0^T}C_0^T Ze^{tA_0}B_0\right\}dt.\end{aligned} \tag{24}$$

In A and B

$$\begin{aligned}\mathcal{H}^{A,B}(Y,Z) &:= \frac{1}{2}\frac{d}{du}\frac{d}{dv}D(A_0+uY,B_0+vZ,C_0)|_{u=v=0}\\ &= \int_0^\infty \rho(t)\,\mathrm{tr}\left\{C_0\left(\int_0^t e^{sA_0}Ye^{(t-s)A_0}ds\right)B_0Z^Te^{tA_0}C_0\right\}dt,\end{aligned} \tag{25}$$

and finally in A and C:

$$\begin{aligned}\mathcal{H}^{A,C}(Y,Z) &:= \frac{1}{2}\frac{d}{du}\frac{d}{dv}D(A_0+uY,B_0,C_0+vZ)|_{u=v=0}\\ &= \int_0^\infty \rho(t)\,\mathrm{tr}\left\{C_0\left(\int_0^t e^{sA_0}Ye^{(t-s)A_0}ds\right)B_0B_0e^{tA_0}Z\right\}dt.\end{aligned} \tag{26}$$