The evolution of phenotypically invariant gene networks with implications for speciation, adaptation, and neutral variation

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

I will outline an analytical theory to study the evolution of biological systems such as gene regulatory networks, borrowing insight and tools from control engineering, systems identification, and dynamical systems theory. I will describe a null model of regulatory network evolution by analytically describing the set of all linear gene networks (of any size) that produce identical phenotypes – and the evolutionary paths connecting them. In the idealized case of a perfectly adapted population, constant selection, and a static environment, we observe neutral evolution as a random walk over the phenotypically- invariant network-space. Under neutral conditions, this model can provide descriptions of expected network size and connectivity under mutation-selection equilibrium, estimate the rate of regulatory rewiring, and the rates at which Dobzhansky-Muller incompatibilities arise in reproductively isolated populations. This analysis provides insight into the mechanisms and parameters important for understanding developmental systems drift, network rewiring, evolvability, epistasis, and speciation, as well as the tenuous connection between network architecture and function.

Introduction

Bridging the gulf between an organism's genome and phenotype is a poorly understood and complex molecular machinery. Progress in a suite of biological subdisciplines is stalled by our general lack of understanding of this molecular machinery: with respect to both its function and evolution. There does exist a growing body of experiment and data on the evolutionary histories and molecular characterizations of particular gene regulatory networks[1;2;3], as well as thoughtful verbal and conceptual models[4;5;6;7]. However, as Hardy and Weinberg taught us over a century ago, verbal theories are often insufficient, if not downright misleading[8;9;10]. This is especially pertinent given the staggering complexity and scope of contemporary research programs. This outlook necessitates the advancement of conceptual frameworks of such precision, only mathematics will suffice. Previously it has been suggested that any idealized study of evolution is incomplete without a mathematically sufficient description of the genotype, phenotype, and transformation from one to the other[11].

The molecular machinery, interacting with the environment, and bridging genotype to phenotype can be mathematically described as a dynamical system – or a system of differential equations[12]. Movement in this direction is ongoing, as researchers have begun to study the evolution of both abstract[13;14;15;16] and empirically inspired computational and mathematical models of gene regulatory networks (GRNs)[17;18;19;20;21;22;23;24]. If we allow the reasonable assumption that the genotype-phenotype map can be represented as a system of differential equations, we can immediately discuss its evolution and function in a much more mechanistic, yet general, manner.

In some fields that seek to fit parametric models to experimental data, such as control theory, chemical engineering, and statistics, it is well known that mathematical models can fundamentally be *unidentifiable* and/or *indistinguishable* – meaning that there can be uncertaintity about an inferred model's parameters or even its claims about causal structure, even with access to complete and perfect data[25;26;27]. Models with different parameter schemes, or even different mechanics can be equally accurate, but still not *actually* agree with what is being modelled. 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[28;29;30]. In chemical engineering, those who study chemical reaction networks sometimes refer to the fundamental unidentifiability of these networks as "the fundamental dogma of chemical kinetics"[31]. In computer science, this

is framed as the relationship among processes that simulate one another[32]. Although this may frustrate the occasional engineer or scientist, viewed from another angle, the concepts of unidentifiability and indistinguishability 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. In fact, evolutionary biologists who study homology and analogy are very familiar with such functional symmetries; macroscopically identical phenotypes in even very closely related species can in fact be divergent at the molecular and sequence level[4;33;34;35;36;37;38].

In this paper we propose a framework to study the evolution of biological systems. To begin, we focus on the evolution of an idealized population. We consider the evolution of a perfectly adapted, large population, evolving in a static environment for an infinite number of generations. Under these ideal circumstances, we expect to observe a "conservation of phenotype," where the population explores the manifold of phenotypically-invariant (or symmetric) genetic and developmental architecutres. We would like to understand which parameters influence the distribution of a population along the manifold of phenotypically invariant genetic systems. Further, we can show how dispersion along this manifold contributes to speciation and evolvability.

The Model I: Gene Networks as Linear Systems

Organisms' phenotypes are constructred by gene by gene by environment interactions. Here we simply define the phenotype to be the organismal temporal molecular dynamics directly under natural selection. The *what*, *when*, and *where*, of an organism's molecules that are physiologically or otherwise relevant for survival. Thus we say that some function $\phi(t)$ is a phenotype where,

$$\phi(t) = \int_0^\infty h(t)u(t)dt, \tag{1}$$

and $h(t)$ is the *impulse response* of the system and $u(t)$ is the *input* function, both functions of time t. The input can be interpreted as the environment, as initial conditions, or otherwise, depending on the biological specifics under study.

Essentially the phenotype $\phi(t)$ is a consequence of an organism's specific gene by gene interactions, given by $h(t)$, reacting further with the local environment, given by $u(t)$.

We describe the impulse response as,

$$h(t) = Ce^{At}B \tag{2}$$

where A is a gene network – a square matrix, and B filters and translates the input to the system. The form of B determines precisely how the state of the external environment influences the internal gene network. C filters and translates the dynamics of the system and precesily determines the output, that is, what is visible to selection.

Generally A can be any real $n \times n$ matrix, B any $n \times \ell$, and C any $\ell \times n$ dimensional matrix. However, for simplicity in exposition (and without loss of generality) we set $\ell = 1$, so that B and C are simply vectors of length n.

Although $\phi(t)$ describes the phenotype given an input, $h(t)$ describes the phenotype subject only to an impulse – an input present initially and absent immediately thereafter. Typically, a system Σ, is defined as,

$$\Sigma = \begin{cases} \dot{\kappa}(t) &= A\kappa(t) + Bu(t) \\ \phi(t) &= C\kappa(t) \end{cases} \tag{3}$$

Variables have the same identities as described above and $\kappa(t)$ is a vector of molecule concentrations at time t. Therefore the molecular concentrations at a specific time are completely determined by the input and gene by gene interactions. Lastly, a portion and/or combination of these molecules, $\phi(t)$, are "observed" by selection (this is in constrast to $\kappa(t)$ – the *kryptotype* – as it is "hidden" from direct selection).

Example 1 (Oscillating Gene Network: Circadian Rhythm)**.**

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$\Sigma = \begin{cases} \dot{\kappa}(t) &= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \kappa(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t) \\ \phi(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \kappa(t) \end{cases}$$

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

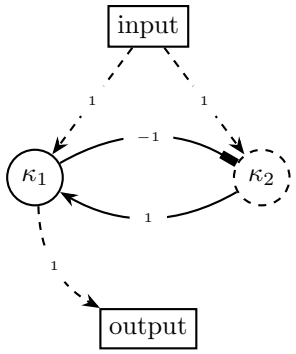
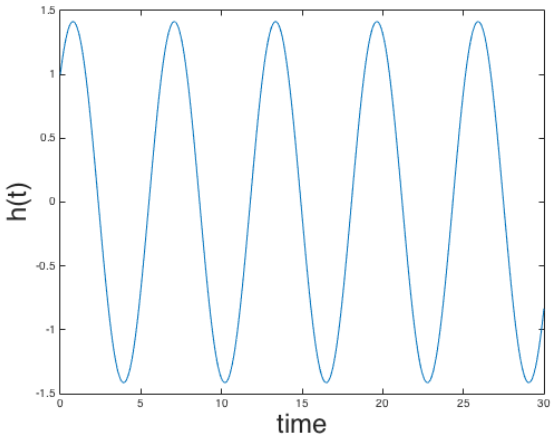


Figure 1: Diagram of Example **??** in the text. *explain what arrows mean if nec*



The Model II: Linear Evolutionary Systems

The literature is filled with detailed observations of molecular systems and their diversity. There are examples of significant diversity in the networks underlying processes such as circadian rhythm[39], cell cycle control[40;41], pattern formation, and metabolism[42;43;44;45;46;47]. Despite a symmetry in functionality or phenotype these systems often differ, sometimes substationaly, at the molecular level. How many different mechanisms have the same function? We urge the reader to consider the metaphorical black box.

Systems with identical external dynamics do not necessarily have identical internal dynamics. Any linear and minimal system – minimal, informally meaning that the system's external dynamics are achieved with the fewest possible number of internal components – has identical external dynamics up to a change of coordinates.

$$\begin{aligned} h(t) &= Ce^{At}B && (4)\\ &= CV^{-1}e^{VAV^{-1}t}VB && (5)\\ &= CV^{-1}Ve^{At}V^{-1}VB && (6)\\ &= Ce^{At}B && (7) \end{aligned}$$

Two systems, $\Sigma = \{A, B, C\}$, and $\bar{\Sigma} = \{\bar{A} = VAV^{-1}, \bar{B} = VB, \bar{C} = CV^{-1}\}$, have the same dynamics if they are related by a change of coordinates.

Although systems may not be identifiable beyond a change of coordinates, at present we are primarily interested in a subset of these systems. That is, systems that not only have equivalent external dynamics, but also equivalent input and output relationships. Formally, this means systems related by a change of coordinates (any invertible matrix V) that leaves B and C invariant:

$$\begin{aligned} VB = B &\implies \bar{B} = B && (8)\\ CV = C &\implies \bar{C} = C && (9) \end{aligned}$$

In other words systems with varying genetic architectures yet identical selection pressures, environment, and phenotype.

Define $V(\tau)$ as the parameterized change of coordinates matrix that preserves B and C, with τ a vector of free parameters. The set of *all* phenotypically invariant (minimal) gene networks is,

$$A(\tau) = V(\tau)A(0)V^{-1}(\tau), \qquad (10)$$

and a *Linear Evolutionary System* is,

$$\Sigma(\tau) = \begin{cases} \dot{\kappa}(t) & = A(\tau)\kappa(t) + Bu(t) \\ \phi(t) & = C\kappa(t) \end{cases} \qquad (11)$$

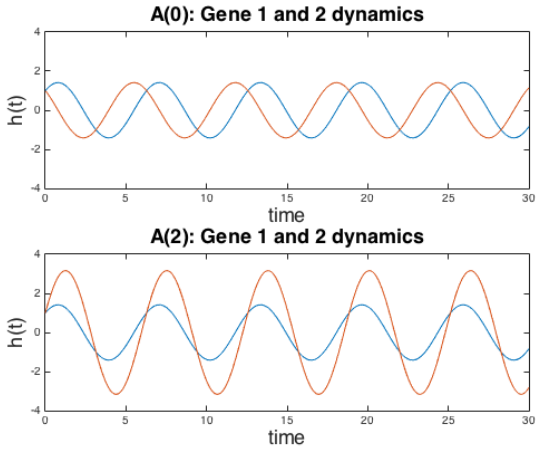
That is, all (linear and minimal) mechanisms capable of producing the same phenotype can be realized by a unique choice of τ in $\Sigma(\tau)$.

More generally, *introduce Kalman*, we denote by $\mathcal{A}_n(A_0)$ the set of all n-dimensional systems equivalent to A_0:

$$\begin{aligned} \mathcal{A}_n(A_0) &= \left\{A : Ce^{At}B = Ce^{A_0 t}B \text{ for } t \geq 0\right\} \\ &= \left\{A : CA^kB = CA_0^kB \text{ for } 1 \leq k \leq n-1\right\}. \end{aligned} \qquad (12)$$

Equivalence of the two characterizations follows from the Cayley-Hamilton theorem. Usually, the dimension n and the reference system A_0 is implicit and we write only $\mathcal{A}$.

Example 2 (External Equivalence does not imply internal equivalence)**.** *Gene 1 dynamics (blue) are equivalent for network architectures $A(0)$ and $A(2)$, however the internal dynamics containing gene 2 (orange) are very different.*



Intraspecific variation and genetic drift

At any given time, there will be a range of network coefficients present in the population due to segregating genetic polymorphism. Over many generations, even if selective pressures do not change, this range of networks will shift as recombination, mutation, and demographic noise create new alleles and shift allele frequencies. How much variation do we expect to find within a population? Is this range limited by available variation or kept in check by selection? How fast will a population explore the space of equivalent networks?

The amount and structure of this standing variation is established over long time scales by many factors, including mutation-selection balance, shifts in the phenotypic optimum, and/or spatial variation in the optimum[?] . Quantitative genetics models of mutation-selection balance predict precise levels and structure of standing variation[? ?] , but it is unclear how well these predictions match reality[?] and how much they are expected to change over time[?] . However, recent empirical work allows us to estimate at least the rough magnitude of variation. XXX FILL IN DETAILS: Mutational scans find this distribution of binding site change per nucleotide change in promoters. Intraspecific variation is often in the range of 0.1%–1%. This suggests typical regulatory binding strengths within a population might vary by up to roughly X%. But, does it? **?** (others?) found evidence that large-effect regulatory mutations are weakly selected against in Drosophila, which suggests that (a) the observed range of variation is somewhat less than this, and (b) the strength of selection on phenotype is sufficient to weakly constrain this variation.

This suggests that within a population, transcription factor binding strengths – the entries of A – vary by around X%, at least for networks whose function is strongly constrained. Subsequent generations for the most part resample from this diversity, so in a population of effective size N_e, simply by the variance of the mean of a random sample, the population mean binding strengths will move a few multiples of $X/\sqrt{N_e}\%$ per generation[?] . (This could be taken as a definition of N_e.) Selection will tend to push this mean towards the optimal set of networks[?] , but mean movement parallel to the optimal set (that leaves the phenotype invariant) is unconstrained unless recombination load is substantial[?] . The action of genetic drift is also strongly determined by covariance between standing genetic variation in different regulatory coefficients – known as the G matrix[?] , covariance which may arise due to functional constraints and/or statistical linkage. There may well be functional constraints – but these are not sufficiently well-known to say anything general about. Linkage will almost certainly lead to covariance if the variation is due to *cis*-regulatory variants, in which case the genetic basis of each *row* of A likely lies within a few kilobases of tightly linked sequence, across which a population may carry only a few common haplotypes. However, covariance due to transiently assembled haplotypes is not expected to be stable over long periods of time – a common *cis*-regulatory haplotype of transcription factor k with particularly strong binding to both i and j (leading to positive covariance between A_{ik} and A_{jk}) is no more likely to appear than one with strong binding to i but particularly weak binding to j (negative covariance). (Such transient covariances may well increase the variance of the per-generation change in network mean, however[?] .) In principle, there may be substantially less variation away from the set of optimal coefficients than there is along the set, due to the action of selection. If so, this might require substantial epistatic load – mortality or reduced fitness of a large proportion of new offspring due to recombination between somewhat incompatible alleles. However, it is

unclear if this is likely to occur and XXX we revisit this below.

It therefore seems reasonable to coarsely model the time evolution of population variation in network coefficents as (a) a "cloud" of width X about the population mean, which (b) moves as an unbiased Brownian motion through the set of network coefficients that give the optimal phenotype. In fact, the population mean will not produce exactly the optimal phenotype, but it will be convenient to refer to this closest point on the optimal set as "the population mean".

Brownian motion on the set of equivalent networks: The set $\mathcal{A}$ is characterized as the solutions to the equations (12), and is hence an algebraic variety. In fact, the Kalman decomposition *XXX above* provides us with an explicit characterization of the set. Above we argued that the population mean set of coefficients A was subject to genetic drift, with each entry A_{ij} changing with mean square displacement $\sigma^2/\sqrt{N_e}$ per generation. Since selection constrains the population mean to stay near to the set of optimal networks $\mathcal{A}$, to a good approximation, the population mean moves as Brownian motion in the space of matrices $\mathbb{R}^{n\times n}$ but conditioned to stay on the optimal set. This Brownian motion on $\mathcal{A}$ is a stochastic process driven by the Laplace-Beltrami operator on $\mathcal{A}$ where this makes sense, with special behavior at the singular points (if any *XXX*). Unlike Brownian motion in flat space, this stochastic process can have *bias* – for instance, it is pushed away from regions of negative intrinsic curvature (because there is "less space" in such regions).

NOT TRUE?: However, the process always has the property that changes in A accumulate at constant (mean squared) rate: if $\mathcal{A}$ is locally isomorphic to $\mathbb{R}^m$ around A_0 (there are m degrees of freedom in the solutions to (12)) then $\mathbb{E}\left[\|A_t - A_0\|^2\right] = m\sigma^2 t + O(t^2)$.

$$A(\tau) \xrightarrow{T} A(\tau + \epsilon) \tag{13}$$

After evolutionary time T, the population's gene network architecture evolves from $A(\tau)$ to $A(\tau+\epsilon)$ with a probability inversely proportional to the magnitude of ϵ and proportional to the magnitude of T. *Change in τ is tracking movement of population mean, not "macro" evolutionary change, as interpreted previously.*

$$\Delta_\tau = \left\| \frac{d}{d\tau} \mathrm{vec}\left[A\left(\tau\right)\right] \right\|^{-1} \tag{14}$$

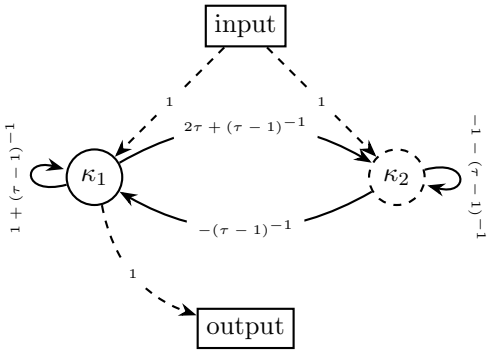
$$A(\tau) \xrightarrow{\mu} A(\tau \pm \mu\Delta_\tau) \tag{15}$$

Example 3 (All Phenotypically Equivalent Oscillators)**.**

$$A(\tau) = \begin{bmatrix} 1 & 0 \\ \tau & 1-\tau \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{\tau}{\tau-1} & \frac{-1}{\tau-1} \end{bmatrix}$$

$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$h(t) = \sin(t) + \cos(t) \qquad \forall\ \tau \neq 1$$



Phenotypic Invariance

First we focus on a simple evolutionary scenario: a large population, perfectly adapted, and in a constant environment. In this circumstance we expect phenotype to be conserved throughout evolutionary time. As such we should only expect the phenotype to change as a consequence of genetic drift (small effective population size), adaptation to new selective and/or environemental pressures. These phenotypic variations should yield distinct signatures. Adaptative changes will change the optimal impulse response function $h(t) \xrightarrow{adaptation} h'(t)$. Genetic drift registers as an increase in the intrapopulation variation in $h(t)$.

Presently we ask two questions, (1) holding $\phi(t)$, $h(t)$, and $u(t)$ constant for evolutionary time T, how much do we expect gene network organization to drift, and (2) does this contribute to speciation, primarily via the fixation of reproductive incompatibilities?

Example 4 (Not all minimal gene networks can drift)**.** *If a gene network is minimal and all the molec-*

ular species involved in the network are under selection, such that C is the $n \times n$ identity matrix ($C = I_n$), the only acceptable change of coordinate matrix is the identify matrix.

$$C \vee B = I \tag{16}$$
$$IV^{-1} = I \tag{17}$$
$$\iff V = I \tag{18}$$

Example 5 (All Non-Minimal Gene Networks can Drift)**.** *Despite the existence of a unique genetic architecture in the minimal case, there still exists an infinite number of systems with larger networks that have identical external dynamics.*

$$h(t) = \widehat{h}(t) \iff \tag{19}$$
$$CA^{\jmath}B = \widehat{C}\widehat{A}^{\jmath}\widehat{C} \qquad \text{for } \jmath = 0, 1, \ldots \tag{20}$$

Any two systems with equivalent impulse responses will have equivalent phenotypes.

add Kalman decomposition stuff here

Speciation via Reproductive Incompatibility

A diploid organism's gene network is simply the average of both of its gene network copies; one from each parent. Further, each haploid parental gene network copy is formed via meiosis – by swapping independent genes randomly. Assuming two distinct but genetically homogeneous populations evolving in allopatry meet and form hybrids, the first generation hybrids (F1s) gene network dynamics will be determined by the average of the parental haplotypes. The second generation hybrids (F2s), however, will be the product of a meiosis between parental haplotypes followed by an averaging of gametes.

Specifically, F_1 is the first generation hybrid gene network architecture formed by mating (averaging) $A(\tau)$ and $A(\hat{\tau})$,

$$F_1(\tau, \hat{\tau}) = \frac{A(\tau) + A(\hat{\tau})}{2}. \tag{21}$$

and F_2 is the second generation hybrid gene network architecture formed by gametes $G(i, \tau, \hat{\tau})$ and $G(j, \tau, \hat{\tau})$. Where each $G(\cdot)$ is formed by randomly swapping rows between $A(\tau)$ and $A(\hat{\tau})$, such that the ith gene comes from $A(\tau)$ (i and j are orthogonal vectors, each element 0 or 1, and $i \neq j$).

$$F_2(i, j) = \frac{G(i, \tau, \hat{\tau}) + G(j, \tau, \hat{\tau})}{2} \tag{22}$$

The fitness of an organism can be computed by comparing its impulse response with the optimal response, *this is either zero or infinite. apply a weighting function (see appendix)?*

$$\mathcal{F}\left(\widehat{h}(t)\right) = \exp\left\{-\int_0^\infty \left\|h(t) - \widehat{h}(t)\right\| dt\right\}. \tag{23}$$

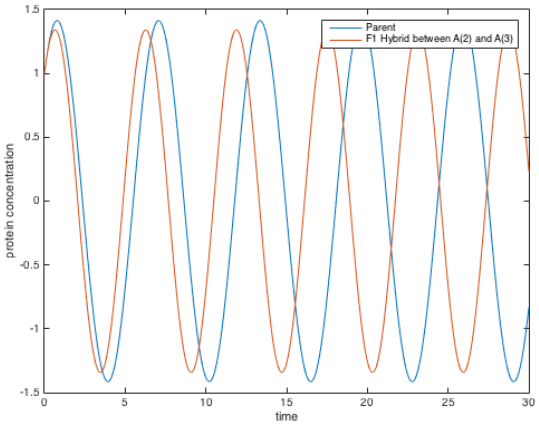
Therefore a hybrid's fitness can be computed by comparing its impulse response with that of its parents.

Example 6 (Quantitative Biological Species Concept)**.** *Two populations are said to be different species if they produce low fitness offspring.*

F1s created by crossing externally equivlaent oscillators $A(0)$ and $A(2)$ have an $h_{F_1}(t) = e^t$, in contrast to both parents with $h_{F_0}(t) = \sin(t) + \cos(t)$. The hybrid phenotype is significantly different (it does not oscillate and increases infinitely) despite the phenotypic equivalence of the parents.

$$\mathcal{F}\left(h_{F_0}\right) = 1$$
$$\mathcal{F}\left(h_{F_1}\right) = 0$$

Example 7 (F1 Reproductive Incompatibility in an Oscillating Gene Network)**.** *DMI examples...*



Example 8 (F2 Reproductive Incompatibility in an Oscillating Gene Network)**.** *F2 versus parental expression dynamics.*

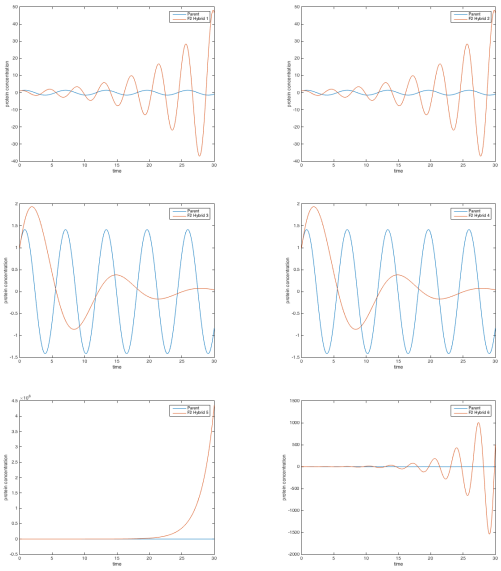


Figure 2: F2s from $A(2)$ and $A(2.1)$.

Example 9 (Not all Networks can Host Incompatibilities)**.** *convex sets cant have DMIs*

$$h(t) = 2e^{-\theta t}$$

Any non-minimal system with rows summing to θ is PI. Further, these systems are closed under averaging (mating) and row swapping (meiosis), leaving all hybrids optimally fit. The set of gene matrices is affine and therefore convex.

Additionl Examples

Metabolic Network

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \text{GAL3} \\ \text{GAL4} \\ \text{GAL80} \\ \text{MIG1} \end{bmatrix}(t) + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \text{galactose} \\ \text{glucose} \\ \text{BTM} \end{bmatrix}(t)$$

$$y(t) = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix} \vec{x}(t)$$

$$H(z) = \begin{bmatrix} 1 \\ z^2 + z + 1 \\ z + 1 \end{bmatrix} \frac{1}{z^3 + z - 1}$$

$$CV = C$$
$$VB = B$$
$$V := \begin{bmatrix} 1 & 0 & a & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, V^{-1} = \begin{bmatrix} 1 & 0 & \frac{-a}{b} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{b} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
$$A(a,b) := VAV^{-1} = \begin{bmatrix} -a & a+1 & \frac{a^2}{b} & -1 \\ 0 & 0 & \frac{-1}{b} & -1 \\ -b & b & a & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The GAL regulatory network in *S. cerevisiae* is modelled here. The activating transcription factor GAL4, which is regulated by the presence of galactose and two other TFs, binds to the promoter region of the GAL regulon – a cluster of genes regulated by the same promoter encoding 3 enzymes (GAL1, GAL7, and GAL10), for the metabolism of the sugar galactose. The repressing transcription factor MIG1, which is activated by the presence of glucose and other proteins omitted here, also binds to the regulon, preventing the expression of the galactose metabolizing enzymes. This network is one of the most studied gene regulatory networks in yeast, and experiments have already demonstrated significant transcriptional variation among different species and genuses of yeast.

In the present model, the A matrix encodes the interactions among GAL3, GAL4, GAL80, and MIG1. B interprets the input: the quantity of glucose, galactose, and the basal transcription (BTM) rates impact of TF concentrations. C represents the promoter region of the GAL regulon and is the sum of the influence of GAL4 (activating) and MIG1 (repressing).

$A(a,b)$ represent alternative network structures that produce identical outputs, with identical input and output transformations – meaning the regulatory contributions of the sugars and the basasl machinery, as well as the promoter of the GAL regulon are held constant. These interactions may be realized in nature, if the regulatory differences predicted by $A(a,b)$ are biologically/mutationally realizable. Mutations influencing the interactions between GAL4 and GAL80 have been experimentally demonstrated[48].

Gap Gene Network

Discussion

mention B part of genetic architecture

Discussion guidelines:

- *Why is this important/useful?*
- *What are the assumptions and shortcomings of the research?*
- *Compare to other studies in the literature.*
- *Future directions.*
- *Wild speculations?*
- *Conclusion and overall impact.*

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, a description of phenomena hitherto only discussed verbally and conceptualy (phenogenetic drift[6], developmental systems drift[4], biological degeneracy[7], *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 offered a novel interpretation of system identification: to see it as an evolutionarily neutral manifold, and not simply a computational nuissance. We have demonstrated a method to analytically determine the set of all phenotypically invarint gene networks; by a simple change of coordinates in the minimal configuration, or more generally by applying the Kalman decomposition in higher dimensions. Further, we emphasize that evolution proceeds through this high dimensional space as stochastic coordinate transformation, constrained by sexual reproduction and selection. 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. We emphasize that these incompatibilities are a consequence of recombining different, yet functionally equivalent, mechanisms. We also suggest that gene networks may not always use their components parsimoniously as network size tends to ratchet up in the absence of strong selection against extra parts. Although unexplored presently, this phenomena may lead to insights on evolvability and developmental innovation. Lastly, we show that hybrid gene networks break down as function of genetic distance, and may, in part, explain broad patterns of reproductive isolation among diverse phyla[49].

As Richard Levins opined, models in population biology face a tradeoff among precision, realism, and generality[50]. As Levins expects, any tractable and general model, such as the present one under discussion, will have limitations. Most notable is linearity. It is often stated that life is not linear. This is often true, however, many of the ideas developed here should be generalizable to nonlinear cases (multilinear systems, say). Futher, we see this as a necessary first step in the direction of more life-like nonlinear evolutionary systems theory. Depending on an actual biological system's particularities, its (potential) nonlinearity, may buffer or exacerbate effects elucidated in this paper, such as the aquisition of Dobzansky-Muller incompatibilities.

This theoretical framework can easily be applied to other interesting questions in evolutionary biology not tackled presently: such as the evolution of linkage, the necessity of network complexity (does evolution tend towards Rube Goldberg or *parsimonious* network organization?), evolvability, structure/function inference, and intrapopulation context dependency of mutational effects, as well as many others.

References

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

[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

[3] 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

[4] John R True and Eric S Haag. Developmental system drift and flexibility in evolutionary trajectories. *Evolution & development*, 3(2):109–119, 2001. 1, 2, 8

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

[6] 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, 8

[7] 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, 8

[8] GH Hardy. Mendelian proportions in a mixed population. *Science*, 28(706):49–50, 1908. 1

[9] Wilhelm Weinberg. Über vererbungsgesetze beim menschen. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, 1(1):440–460, 1908. 1

[10] 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

[11] Richard C Lewontin et al. *The genetic basis of evolutionary change*, volume 560, chapter The Structure of Evolutionary Genetics, pages 12–16. Columbia University Press New York, 1974. 1

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

[13] 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

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

[15] 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

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

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

[18] 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

[19] 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

[20] 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

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

[22] 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

[23] 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

[24] 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. 1

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

[26] M Grewal and K Glover. Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on automatic control*, 21(6):833–837, Dec 1976. 1

[27] 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

[28] Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *J.S.I.A.M Control*, 1963. 1

[29] 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. 1

[30] Lotfi A Zadeh and Charles A Deoser. *Linear system theory*. Robert E. Krieger Publishing Company Huntington, 1976. 1

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

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

[33] 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

[34] 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

[35] Jeff Vierstra, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R Scott Hansen, Sandra Stehling-Sun, Peter J Sabo, Rachel Byron, Richard Humbert, et al. Mouse regulatory dna landscapes reveal global principles of cis-regulatory evolution. *Science*, 346(6212):1007–1012, 2014. 2

[36] Andrew B Stergachis, Shane Neph, Richard Sandstrom, Eric Haugen, Alex P Reynolds, Miaohua Zhang, Rachel Byron, Theresa Canfield, Sandra Stelhing-Sun, Kristen Lee, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–370, 2014. 2

[37] Matthew B Taylor, Joann Phan, Jonathan T Lee, Madelyn McCadden, and Ian M Ehrenreich. Diverse genetic architectures lead to the same cryptic phenotype in a yeast cross. *Nature communications*, 7, 2016. 2

[38] Takeshi Matsui, Robert Linder, Joann Phan, Fabian Seidl, and Ian M Ehrenreich. Regulatory rewiring in a cross causes extensive genetic heterogeneity. *Genetics*, 201(2):769–777, 2015. 2

[39] Aziz Sancar. The intelligent clock and the rube goldberg clock. *Nature structural & molecular biology*, 15(1):23–24, 2008. 3

[40] 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. 3

[41] Stephen E Kearsey and Sue Cotterill. Enigmatic variations: divergent modes of regulating eukaryotic dna replication. *Molecular cell*, 12(5):1067–1075, 2003. 3

[42] 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. 3

[43] 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. 3

[44] 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. 3

[45] 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. 3

[46] 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. 3

[47] 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. 3

[48] Yan Li, Guanjun Chen, and Weifeng Liu. Alterations in the interaction between gal4 and gal80 effect regulation of the yeast gal regulon mediated by the f box protein dsg1. *Current microbiology*, 61(3):210–216, 2010. 7

[49] 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. 8

[50] Richard Levins. The strategy of model building in population biology. *American scientist*, 54(4):421–431, 1966. 8

A Differentiating the fitness function

In case this is useful...

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:

$$\begin{aligned} D(A) &:= \int_0^\infty \rho(t) \left| Ce^{At}B - Ce^{A_0 t}B \right|^2 dt \\ &= \int_0^\infty \rho(t) \left| C\left(e^{At} - e^{A_0 t}\right)B \right|^2 dt \\ &= \int_0^\infty \rho(t) C\left(e^{At} - e^{A_0 t}\right) BB^T \left(e^{At} - e^{A_0 t}\right)^T C^T dt \end{aligned} \tag{24}$$

How does this change with A? Since

$$\frac{d}{du} e^{(A+uZ)t}|_{u=0} = \int_0^t e^{As} Z e^{A(t-s)} ds, \tag{25}$$

we have that

$$\begin{aligned} \frac{d}{du} D(A+uZ)|_{u=0} &= 2\int_0^\infty \rho(t) C \left(\int_0^t e^{As} Z e^{A(t-s)} ds \right) BB^T \left(e^{At} - e^{A_0 t}\right)^T C^T dt \\ &= 2\int_0^\infty \int_0^t \rho(t) C e^{As} Z e^{A(t-s)} BB^T \left(e^{At} - e^{A_0 t}\right)^T C^T ds dt. \end{aligned} \tag{26}$$