The Evolution of Phenotypically Invariant Genetic Networks

Joshua S. Schiffman and Peter L. Ralph

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 [???], as well as thoughtful verbal and conceptual models [??]. However, as Hardy and Weinberg taught us over a century ago, verbal theories are often insufficient, if not downright misleading [???]. 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 [?].

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[?]. Movement in this direction is ongoing, as researchers have begun to study the evolution of both abstract [????] and empirically inspired computational and mathematical models of gene regulatory networks (GRNs) [????????]. 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 [???]. 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 [???]. 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" [?]. In computer science, this is framed as the relationship among processes that simulate one another [?]. Although this may frusturate 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 [**???????**].

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 complex gene by gene and gene by environment interactions. Although a very simplified description, here we define the phenotype to be the spatio-temporal molecular dynamics directly under natural selection. Basically the *what*, *when*, and *where*, of an organism's molecules that are physiologically or otherwise important and relevant to survival. Thus we say that some function $f(t)$ is a phenotype where,

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

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

Essentially the phenotype $f(t)$ 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.

Above, t refers to time. Although $f(t)$ 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 defines as,

$$\Sigma = \begin{cases} \dot{x}(t) & = Ax(t) + Bu(t) \\ y(t) & = Cx(t) \end{cases} \tag{3}$$

Variables have the same identities as described above and $x(t)$ is a vector of molecule concentrations at time t. Therefore the molecular concentrations at a specific time are completely determined by the environmental input and gene by gene interactions. Lastly, a portion and/or combination of these molecules, $y(t)$, are "observed" by selection.

The Model II: Linear Evolutionary Systems

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 up to 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 equivlanet input relationships and under equivalent selective pressures. Formally, systems related by a change of coordinates that leave B

and C invariant:

$$VB = B \implies \bar{B} = B \tag{8}$$
$$CV = C \implies \bar{C} = C \tag{9}$$

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

Define $V(\tau)$ as the 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,

$$A(\tau) = V(\tau)A(\tau_0)V^{-1}(\tau), \tag{10}$$

and a Linear Evolutionary System is,

$$\Sigma(\tau) = \begin{cases} \dot{x}(t) & = A(\tau)x(t) + Bu(t) \\ y(t) & = Cx(t) \end{cases} \tag{11}$$

Evolution thus proceeds as a random walk in phenotypically invariant network space.

$$A(\tau) \xrightarrow{T} A(\tau + \epsilon) \tag{12}$$

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.

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 $f(t)$, $h(t)$, 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?

Not all minimal gene networks can drift

If a gene network is minimal and all the molecular 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.

$$C \vee B = I \tag{13}$$
$$IV^{-1} = I \tag{14}$$
$$\iff V = I \tag{15}$$

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.

$$h(t) = \widehat{h}(t) \iff \tag{16}$$
$$CA^kB = \widehat{C}\widehat{A}^k\widehat{C} \qquad \text{for } k = 0, 1, \ldots \tag{17}$$

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

Speciation via Reproductive Incompatibility

Define reproduction in diploids as first, the recombination of unlinked genes to make gametes, and second, as the averaging of two individual parental gametes to produce an offspring. Assuming parental populations are both phenotypically identical and genetically homogenous within each population, first generation hybrids (F1s) can be computed by averaging the two parental gene networks. Second generation hybrids (F2s) can be computed by first swapping the genes between the two parental gene networks, and next averaging these hybrid gametes.

$$A_{(F1)}(\tau_i, \tau_j) = \frac{1}{2}\left(A(\tau_i) + A(\tau_j)\right)$$
$$h_{(F1)}(t) = Ce^{\frac{t}{2}(A(\tau_i) + A(\tau_j))}B$$

$$G_{(F2)}(r) = Q(r)A(\tau_i) + \left(I - Q(r)\right)A(\tau_j)$$

$$h_{(F2,r,r')}(t) = Ce^{\frac{t}{2}\left(G(r) + G(r')\right)}B$$

$$\Phi(h'(t)) = e^{-\int_0^\infty \left\| h(t) - h'(t) \right\| dt}$$

Examples

Cell Cycle Control or Circadian Rhytm

Metabolic Network

Gap Gene Network

Discussion