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)**.**

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

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

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

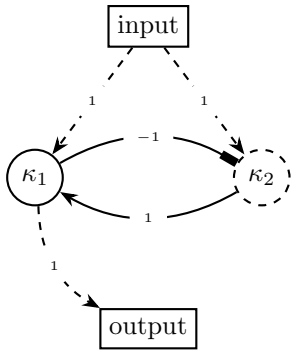
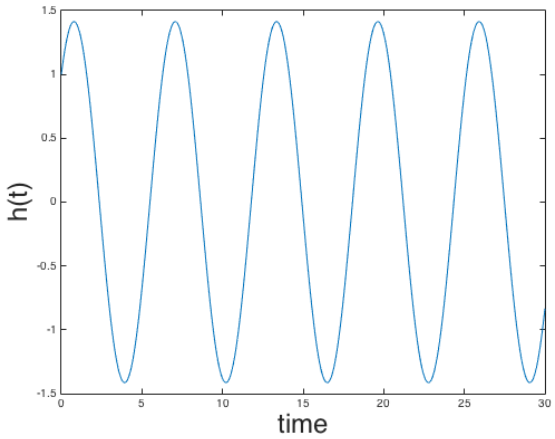


Figure 1: Diagram of Example **??** in the text. *explain what arrows mean if nec*



The Model II: Linear Evolutionary Systems

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.

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

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:

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

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

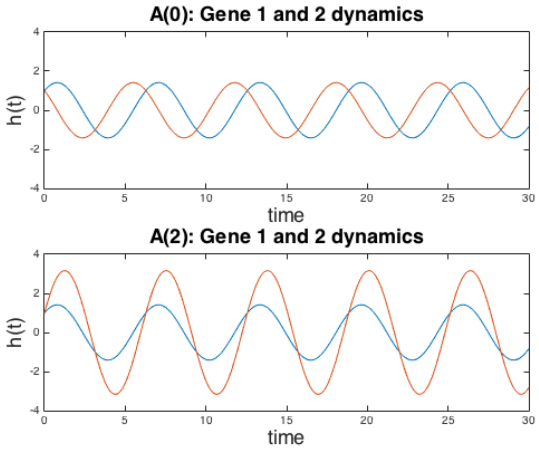
Define $V(\tau)$ as the parameterized change of coordinates matrix that preserves B and C, with τ a vector of free parameters. The set of *all* phenotypically invariant (minimal) gene networks is,

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

and a *Linear Evolutionary System* is,

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

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.*



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. *Change in τ is tracking movement of population mean, not "macro" evolutionary change, as interpreted previously.*

$$\Delta_\tau = \left\| \frac{d}{d\tau} \mathrm{vec}\left[A\left(\tau\right)\right] \right\|^{-1} \tag{13}$$

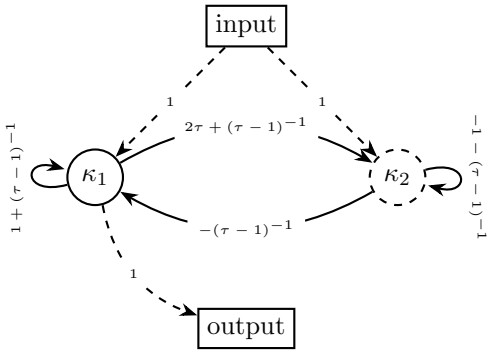
$$A(\tau) \xrightarrow{\mu} A(\tau \pm \mu\Delta_\tau) \tag{14}$$

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 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.*

$$C \vee B = I \tag{15}$$

$$IV^{-1} = I \tag{16}$$

$$\iff V = I \tag{17}$$

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.*

$$h(t) = \widehat{h}(t) \iff \tag{18}$$

$$CA^\jmath B = \widehat{C}\widehat{A}^\jmath\widehat{C} \qquad \textit{for } \jmath = 0, 1, \ldots \tag{19}$$

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 $A(\hat{\tau})$,

$$F_1(\tau, \hat{\tau}) = \frac{A(\tau) + A(\hat{\tau})}{2}. \tag{20}$$

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 $i \neq j$).

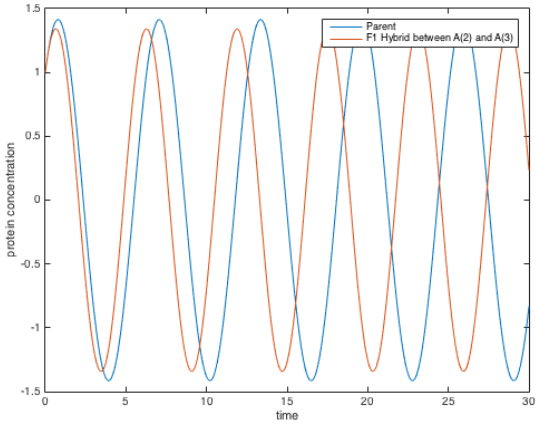
$$F_2(i, j) = \frac{G(i, \tau, \hat{\tau}) + G(j, \tau, \hat{\tau})}{2} \tag{21}$$

The fitness of an organism can be computed by comparing its impulse response with the optimal response,

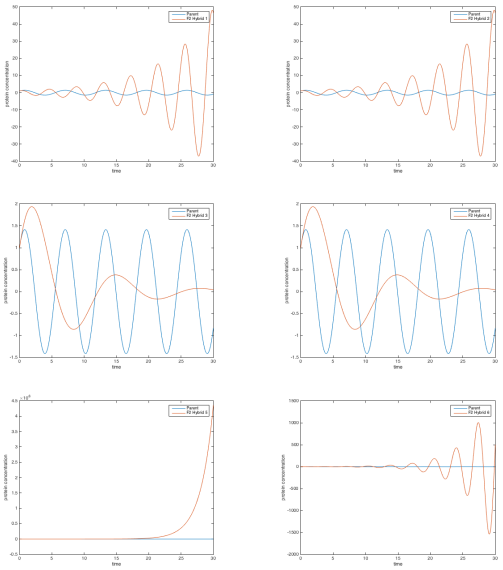
$$\mathcal{F}\left(\widehat{h}(t)\right) = \exp\left\{-\int_0^\infty \left\|h(t) - \widehat{h}(t)\right\| dt\right\}. \tag{22}$$

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

Example 6 (F1 Reproductive Incompatibility in an Oscillating Gene Network)**.** *DMI examples...*



Example 7 (F2 Reproductive Incompatibility in an Oscillating Gene Network)**.** *F2 versus parental expression dynamics.*



Example 8 (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

Gap Gene Network

Discussion

mention B part of genetic architecture

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

[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

[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] 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