Chance and necessity in gene network complexity

Joshua S. Schiffman[†] Peter L. Ralph[†‡]

[†]Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, U.S.A.
[‡]Departments of Mathematics and Biology & The Institute for Ecology and Evolution, University of Oregon, Eugene, Oregon 97403, U.S.A.

jsschiff@usc.edu plr@uoregon.edu

July 19, 2018

Abstract

How does the balance of selection and drift shape gene regulatory network architecture? Do networks tend to ratchet up in complexity?

1 Introduction

It is an outstanding problem in biology to identify the myriad of interactions among genes and their products and further to ascertain the functional and adaptive significance of these interations. At first glance, through the incomplete application of evolutionary principles, one might draw the conclusion that the architecture of genetic interactions within a cell are precisely so to carry out adaptive functions. It has, however, been long known and demonstrated that this hypothesis fails under scrutiny; population genetic models generally do not support such a view Lynch [2007].

It has been suggested that the architecture of biological systems can also largely be influenced by non-adaptive or neutral processes. This is intuitively appealing to its proponents as many systems seem to be of "baroque" design, unnecessarily complex, or in other ways unneccesarilly complicated, such as the Rube Goldberg structure of eukaryotic circadian rhythms [Sancar, 2008]. Along a similar vein, the apparent "gratuitous complexity" of RNA editing [Covello and Gray, 1993] and the eukaryotic splicesome [Nilsen, 2003, Gray et al., 2010] may be the consequence of non-adaptive evolutionary processes, such as a neutral complexity ratchet, sometimes refered to as *constructive neutral evolution* [Stoltzfus, 1999, 2012, Lukeš et al., 2011]. This ratchet proncess, as previously pointed out [Gray et al., 2010] is conceptually similar to *contingent irreversibility* [Szathmáry and Smith, 1995] – the "accidental" process by which mitchondria may have lost the ability to independently replicate, after symbiotically residing within a larger cell, and to the duplication-degeneration-complementation process by which redudant duplicate genes evolve away from expendability [Lynch and Force, 2000].

Although the invocation of neutral processes is appealing, and an attractive alternative, to what is sometimes called the pan-adaptationist view, it is important to understand the balance of both neutral and adaptive processes. The structure of these interactions, or genetic networks, is the consequence of overlapping forces: selection for adaptive function, population factors and genetic drift. In this paper we apply tools from system theory to decompose the affects of these forces on network structure. Here we aim to clarify the interplay of these processes by using explicit (mathematical and computational) methods, as the current models describing the evolution of biological complexity are largely verbal and intuitive.

Within the framework of system theory, we can explicitly interpret the function of a genetic network and also determine the parsimony of such a network. That is, given some adaptive function – e.g. the oscillating transcription of a gene regulatory network involved in circadian rhythm – we can explicitly evaluate a particular network's fitness for such a function and further ask how parsimonious such a network is. The

parsimony of a system is a measure of the number of interacting genes within the network. If a specific function, such as an oscillator, can be achieved, in principle, by as few as two genes, then any such network involving more than two genes can be viewed as (unncessarily) complex, at least with respect to its imagined function.

Within this framework we aimed to understand how the balance of evolutionary forces influence network architecture. Under which circumstances, if any, do we observe highly parsimonious versus highly complex networks?

2 Results

System theory As in Schiffman and Ralph [2018], we model gene regulatory networks as linear dynamical systems.

We use a model of gene regulatory networks that describes the temporal dynamics of a collection of n coregulating molecules – such as transcription factors – as well as external or environmental inputs. We write $\kappa(t)$ for the vector of n molecular concentrations at time t. The vector of m "inputs" determined exogenously to the system is denoted $u(t)$, and the vector of ℓ "outputs" is denoted $\phi(t)$. The output is merely a linear function of the internal state: $\phi_i(t) = \sum_j C_{ij}\kappa_j(t)$ 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 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 any $n \times m$, and C any $\ell \times n$ dimensional matrix, with usually ℓ and m less than n. We think of the system as the triple (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 $h(t) := Ce^{At}B$.

In terms of gene regulatory networks, 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 i^{th} 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 determine the rate of production of downstream enzymes.

System drift and the Kalman decomposition As we previously demonstrated in Schiffman and Ralph [2018], system drift can be modeled within this framework by either applying a simple coordinate transformation to a minimal system realization, or more generally, by applying the Kalman Decomposition. The Kalman Decomposition thus characterizes the set of all phenotypically equivalent linear gene regulatory networks, of any finite network size.

This characterization immediatley leads us to ask, what determins network size and complexity?

To address this question, we first give a precise definition of "complexity", as used in this article. Here, we refer to a system as *complex* if it is non-minimal. Thus, any system that could, in principle, successfully perform its function, exactly, with fewer genes, is deemed to be complex.

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

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

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}).

This decomposition is unique up to a change of basis that preserves the block structure. In particular, the minimal subsystem obtained by the Kalman decomposition is unique up to a change of coordinates. This implies that there is no equivalent system with a smaller number of kryptotypic dimensions than the dimension of the minimal system. It is remarkable 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. (However, some of these subspaces may affect how the system deals with noise.)

Fitness function We will quantify 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, $h_0(t)$ is the *optimal* impulse response function and define the "distance to optimum" of another impulse response function to be

$$D(h) = \left(\int_0^\infty \rho(t) \|h(t) - h_0(t)\|^2 dt \right)^{1/2}. \tag{3}$$

The fitness of an organism is then given by,

$$\exp\left(-\frac{D(h)}{\gamma} \right) \tag{4}$$

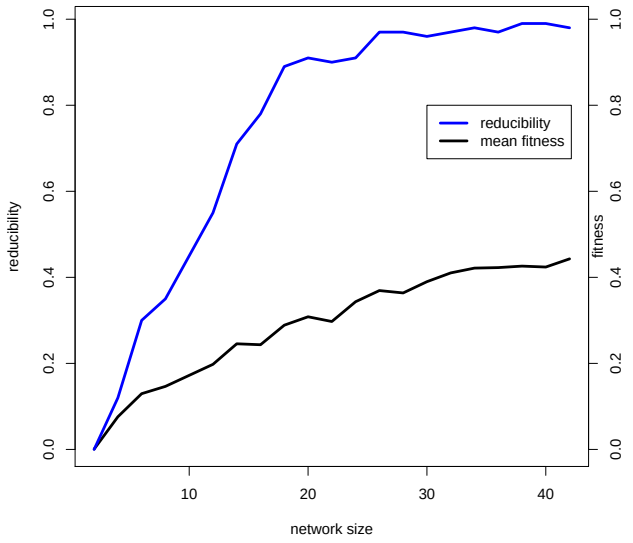
Ratchet Propensity To assess the possibility of a gene regulatory network ratchet we asked whether or not non-minimal gene regulatory neworks were robust to gene deletions. We reasoned that, although the removal of an initially superfluous gene may be incosequential, that following a period of system drift the initially superfluous gene may be difficult, or even impossible to remove without compromising network function, and thus organismal fitness.

If such a process were to occur, network complexity may *ratchet* up in complexity over evolutionary time, as more and more genes become essential for proper network functioning. Furthermore, if this process were to go on indefinitely, perhaps as more and more genes contribute to network functioning, the phenotypic

impact of removing a gene may diminish with network size, such that networks can more easily be pruned. If so, there may be some equilibrium network size, where a system ratchets up in complexity, up to the point at which gene deletions become easily tolerated.

To examine this process, we randomly generated phenotypically equivalent systems of varying network sizes, sequentially deleting genes (with replacement), and then subsequentially measuring fitnesses. If a system could withstand the removal of at least one gene, we say it is reducibile.

PLOT OF REDUCIBILITY



Simulation Results In the previous seciton, we plotted network reducibility versus network size. This plot suggests that non-minimal networks, on average, are less reducibile than larger networks. Further, it appears that as networks get larger, reducibility tends towards 1.

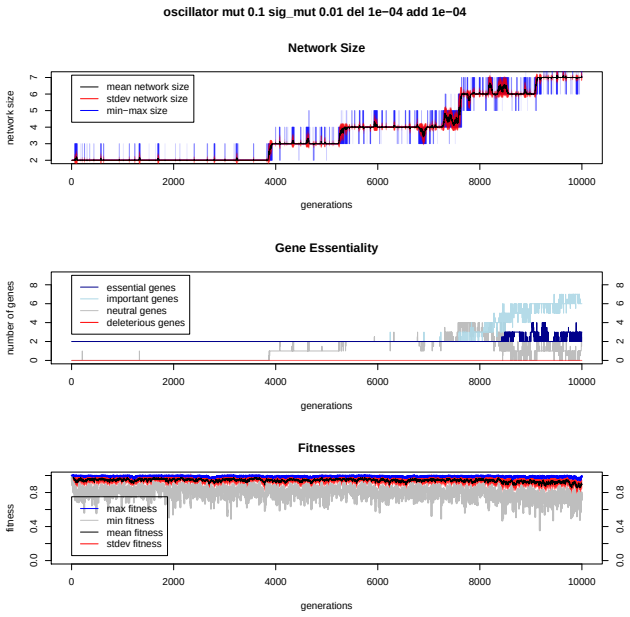
Despite this demonstration, it is not clear that a population evolving under stabilizing selection will ratchet up in complexity. To assess this potential we simulated the evolution of a population under stabilizing selection.

We simulate the evolution of a population of 100 individuals for $10,000$ generations. We vary gene mutation, addition and deletion rates, as well as the average mutational magnitude – that is the magnitude of the change to a system regulatory coefficient. We can also choose any linear dynamical system, as well as the initial network size, whether its minimal or non-minimal at generation zero.

PLOTS OF EVOLUTION SIMULATIONS FOR DIFFERENT PARAMETER SCHEMES AND INITIAL CONDITIONS

Depending on the parameter scheme and the initial system conditions chosen, we observe very different evolutionary outcomes. If we run 80 simulations, starting with a minimal, two-dimensional oscillator, varying each parameter by an order of magnitude ranging in values from 0.1 to 0.0001 such that gene addition rates are always less than or equal to deletion rates, and where deletion rates are always less than or equal to mutation rates, we do not observe any network growth, except for one case, when deletion and addition rates are not equivalent. Further, we only observe network growth in a subset of these simulation, where both the mutation rate and mutational magnitude are relatively high, either 0.1 or 0.01.

Upon further analyzing the simulation results in which systems tend to ratchet up in complexity, we also see that under some parameter schemes, this apparent complexity is often trivial – that is, the excess genes present in the system do not contribute significantly to network function, and can easily be pruned without affecting phenotype or fitness.

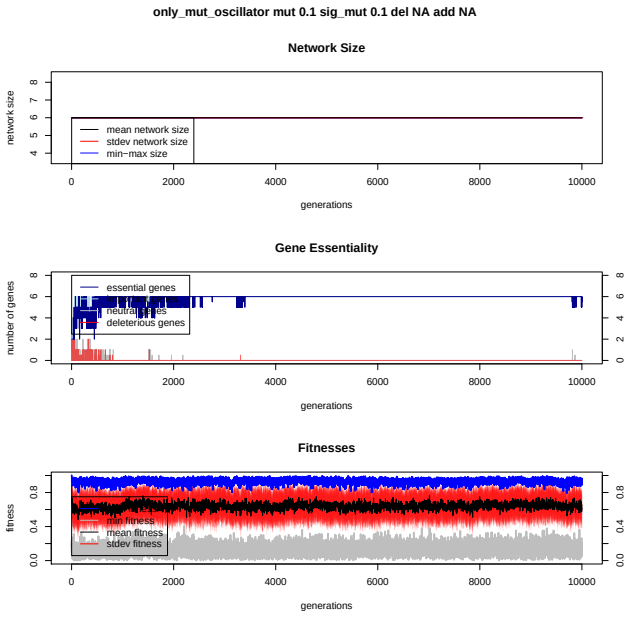


PLOTS WITH GENE ESSENTIALITY COLORED IN

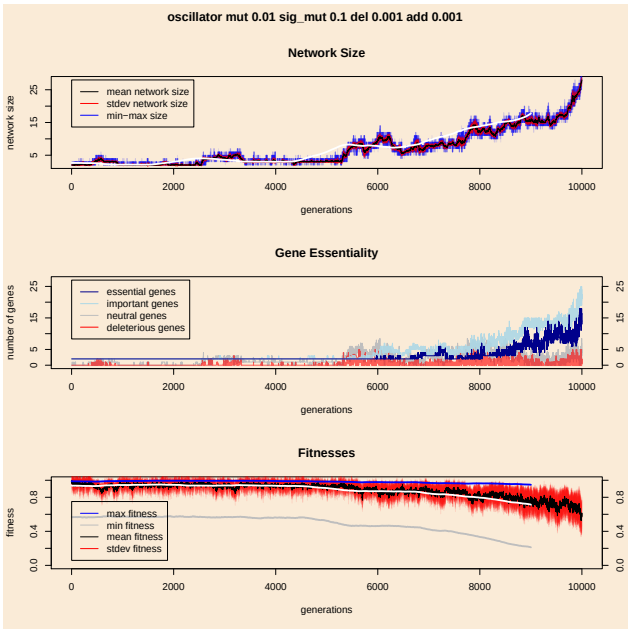
Given that some paramer schemes lead to highly complex gene regulatory networks, whereas others lead to simple, streamlined network architectures, we sought to determine the equilibrium network sizes, given evolutionary parameters.

Simulations without gene addition and deletion dynamics Depending on the biological details of system evolution gene addition and deletion as modeled above may be a fairly inaccurate mechanism. The mechanism, first adds or removes empty gene "slots" which can subsequntially mutate to interact neutrally, deleteriously, or beneficially with the current system. In some cases, genes may not necessarily join functional networks in such a manner, and may instead be recruited from other contemporaneously active regulatory networks. If so, a more accurate model might not allow a mechanism to remove or add new gene slots, but more simply, use a mechanism that recruits already available and active genes. In this case, we may start with a minimum system with a fixed size n', of which only a subset of genes, say the minimum number $n_{\min}$, are initially active (i.e. non-zero), and $n' - n_{\min}$ inactive (i.e. set to zero) genes. In such a scenario, our simulations show that these networks tend to slowly recruit all possible available genes, such that after sufficient time, most, if not all of the n' genes are functionall important to network function, and consequential if removed. This seems to be just a general case of system drift.

PLOT OF STATIC NETWORK SIZE



Mutational Meltdown PLOTS and analysis of when mutational meltdown occurs



Moonlighting among systems If two functionally independent regulatory systems are active within an organism, such that an organism's fitness is a function of two phenotypic traits, given by the systems

(A, B, C) and $(\widehat{A}, \widehat{B}, \widehat{C})$, with state spaces given by,

$$\begin{aligned} \dot{\kappa}(t) &= A\kappa(t) + Bu(t) \\ \phi(t) &= C\kappa(t) \end{aligned} \qquad \begin{aligned} \dot{\hat{\kappa}}(t) &= \widehat{A}\hat{\kappa}(t) + \widehat{B}\hat{u}(t) \\ \hat{\phi}(t) &= \widehat{C}\hat{\kappa}(t). \end{aligned} \tag{5}$$

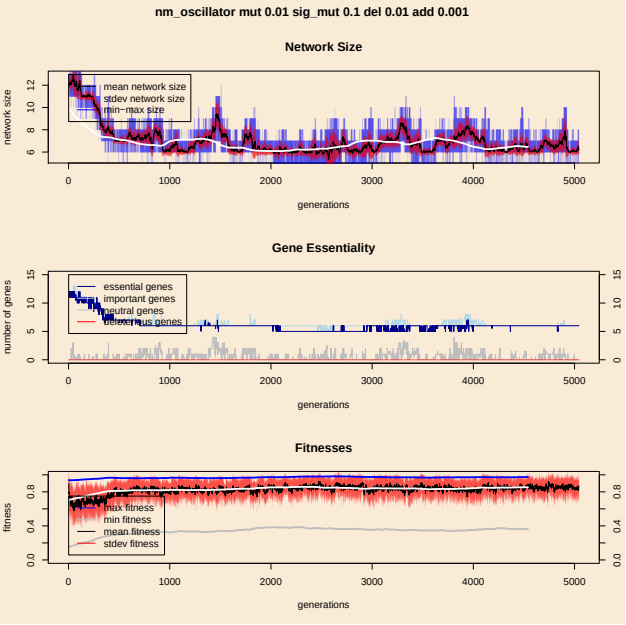
We can represent this as one composite system,

$$\begin{aligned} \begin{bmatrix} \dot{\kappa}(t) \\ \dot{\hat{\kappa}}(t) \end{bmatrix} &= \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} \begin{bmatrix} \kappa(t) \\ \hat{\kappa}(t) \end{bmatrix} + \begin{bmatrix} B & 0 \\ 0 & \widehat{B} \end{bmatrix} \begin{bmatrix} u(t) \\ \hat{u}(t) \end{bmatrix} \\ \begin{bmatrix} \phi(t) \\ \hat{\phi}(t) \end{bmatrix} &= \begin{bmatrix} C & 0 \\ 0 & \widehat{C} \end{bmatrix} \begin{bmatrix} \kappa(t) \\ \hat{\kappa}(t) \end{bmatrix} \end{aligned} \tag{6}$$

If we then apply a coordinate transformation to this composite system, we will preserve its overall function and phenotype, however, interconnections can form between the systems, and the systems may become intertwined by genetic moonlighting. As such, the phenotypically equivalent space for the composite system includes network organizations with interconnections between the systems. In fact, the vast majority of network realizations following a coordinate transformation will include non-zero terms in the off-diagonal entries of the the composite matrices. This is just a special case of system drift. In this case complexity, as defined as superfluous genes (with respect to both systems), does not increase, however the complexity with respect to each individual system does increase. This model seems to predict that pleiotropy may tend to increase among systems simultaneously active within a cell.

This may be an illustration of what is observed in Smith and Keeling [2015].

Non-minimal Start PLOTS where simulation starts at a non-minimal state – selection can reduce system size, but can get "stuck" in a non-minimal spot.



Equilibrium Network sizes If we imagine a strictly neutral gene addition and deletion scenario such that there is no mutational target size penalty, we might expect networks to reach an equillibrium size where its reducibility is near 1. However, in this framework, we also know that the mutational target size of a network will increase quadratically with the number of genes involved, and depending on the typical costs of mutation s and deletion s^{DEL}, selection may oppose gene addition. To quantify the balance of these forces we derive the following equation; network size should be determined by a balance of both neutral and selective evolutionary forces as approximated by,

$$\mathbb{E}\left[n_{t+1}\right] \approx n_t \left(1 + \frac{p_{\mathrm{add}}\left(1 - p_{\mathrm{mut}} 4 n s_n\right) - p_{\mathrm{del}}\left(1 - s_n^{\mathrm{DEL}}\right)}{\bar{w}}\right). \tag{7}$$

This equation estimates the expected network size in the following generation n_{t+1}. This is a consequence of the current network size, the mutation rate, the addition and deletion rates, as well as the average fitness costs of mutation and gene deletions.

This equation can be derived by considering the following table, such that mean network size is computed as,

$$\frac{\sum_{\mathrm{sizes}}(\%\mathrm{size}) \times (\text{fitness at size}) \times \mathrm{size})}{\sum_{\mathrm{sizes}}(\%\mathrm{size}) \times (\text{fitness at size})} \tag{8}$$

Offspring size	%	Mean fitness relative to parent
$n-1$	np_{del}	$1 - s_n^{\mathrm{DEL}}$
n	$1 - (p_{\mathrm{del}} + p_{\mathrm{add}})n$	$1 - p_{\mathrm{mut}} n^2 s_n$
$n+1$	np_{add}	$1 - p_{\mathrm{mut}}(n+1)^2 s_{n+1}$

3 Discussion

We show that similar to Stoltzfus' tangling wires analogy [Stoltzfus, 2012], gene regulatory networks will tend to incorporate all available genes over evolutionary time. This process is essentially an entropic one, as the function of the network slowly diffuses across the available, actively transcribed genes – if a gene can get involved, it likely will get involved. However, it is much less likely that new genes will be added to a network if they are not already, somehow functionally, or at least presently, active. Only under sufficiently high (although perhaps biologically realistic) gene addition and mutation rates, will networks ratchet up in complexity, by adding *de novo* components. Continuing with Stolzfus' analogy, this is the difference between simply scrambling already present wires versus adding and then scrambling wires. In the latter case, if scrambling does not occur rapidly enough, selection may remove the superfluous wires, as they increase the risk of malfunction; in biological terms, these extra genes increase, however slightly, the mutational target size of the network. There is another important distinction between these two processes: there is no need to invoke a ratchet like mechanism to explain the scrambling-only process (also referred to as *moonlighting* [Speijer, 2011]) – whether or not a scrambled network can function with or without one of its components is not the force that adds complexity to networks – it is simply due to the size of the neutral, phenotypically-equivalent, network space. If the neutral network space is sufficiently large, the probability that a system finds itself organized in the simplest (or nearly simplest) configuration becomes exceedingly improbable. In this view, this process is just a special case of *system drift* [True and Haag, 2001, Schiffman and Ralph, 2018]. Furthermore, our results suggest that if a gene regulatory network is recruited to perform a new or modified function (due to changes in selective pressures), upon reorganizing, the network may be unable to reduce its "excess" complexity sufficiently, as it may find itself mutationally disconnected from simpler network organizations. It is important to note, however, that these results are specific to our modeling framework of regulatory networks, and may not translate to other discussions of complexity, such as the evolution of the spliceosome.

References

PatrickS Covello and MichaelW Gray. On the evolution of rna editing. *Trends in Genetics*, 9(8):265–268, 1993. 1

Michael W Gray, Julius Lukeš, John M Archibald, Patrick J Keeling, and W Ford Doolittle. Irremediable complexity? *Science*, 330(6006):920–921, 2010. 1

Julius Lukeš, John M Archibald, Patrick J Keeling, W Ford Doolittle, and Michael W Gray. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB life*, 63(7):528–537, 2011. 1

Michael Lynch. The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8 (10):803–813, 2007. 1

Michael Lynch and Allan Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000. 1

Timothy W Nilsen. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25 (12):1147–1149, 2003. 1

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

J.S. Schiffman and P.L. Ralph. System drift and speciation. *bioRxiv*, page 231209, 2018. 2, 8

David Roy Smith and Patrick J Keeling. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences*, page 201422049, 2015.

Dave Speijer. Does constructive neutral evolution play an important role in the origin of cellular complexity? making sense of the origins and uses of biological complexity. *Bioessays*, 33(5):344–349, 2011. 8

Arlin Stoltzfus. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2): 169–181, 1999. 1

Arlin Stoltzfus. Constructive neutral evolution: exploring evolutionary theorys curious disconnect. *Biology direct*, 7(1):35, 2012. 1, 8

Eörs Szathmáry and John Maynard Smith. The major evolutionary transitions. *Nature*, 374(6519):227–232, 1995. 1

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