

# Model of haplotype and phenotype in the evolution of a duplicated autoregulatory activator

Srinandan Dasmahapatra  
 Electronics and Computer Science  
 Institute for Life Sciences  
 Faculty of Physical and Applied Sciences  
 University of Southampton, SO17 1BJ, UK  
*email:sd@ecs.soton.ac.uk*

---

## Abstract

Gene duplication is believed to play a major role in the evolution of genomic complexity. The presence of a duplicate frees a gene from the constraint of natural selection, leading to its loss of function or the gain of a novel one. Alternately, a pleiotropic gene might partition its functions among its duplicates, thus preserving both copies. Such arguments invoking duplication for novelty or specialisation are not true of diploid genotypes, but only of haplotypes. In this paper, we study the consequences of regulatory interactions in diploid genotypes and explore how the context of allelic interactions gives rise to dynamical phenotypes that enable duplicate genes to spread in a population. The regulatory network we study is that of a single autoregulatory activator gene, and the two copies of the gene diverge either as alleles in a diploid species or as duplicates in haploids. These differences are in their transcriptional ability – either via alterations to its activating domain, or to its *cis*-regulatory binding repertoire. When *cis*-regulatory changes are introduced that partition multiple regulatory triggers among the duplicates, it is shown that mutually exclusive expression states of the duplicates that emerge are accompanied by a back-up facility: when a highly expressed gene is deleted, the previously unexpressed duplicate copy compensates for it. The diploid version of the regulatory network model can account for allele-specific expression variants, and a model of inheritance of the haplotype network enables us to trace the evolutionary consequence of heterozygous phenotypes. This is modelled for the variations in the activating domain of one copy, whereby stable as well as transiently bursting oscillations ensue in single cells. The evolutionary model shows that these phenotypic states accessible to a diploid, heterozygous genotype enable the

spread of a duplicated haplotype.

---

## 1. Introduction

Gene duplication is a major source of genomic expansion and is believed to underlie the evolution of complex biological functions [1; 2]. Functions encoded by two copies of the same gene are redundant making the loss of function of a duplicate by mutation a likely outcome [3]. Hence, the observed abundance of duplicates in plant and animal genomes makes the retention of duplicate genes a much studied problem. The loss of selection pressure on the duplicate can be viewed not as a prerequisite for its elimination, but as an opportunity for it to become more abundant by acquiring a new and fitness-enhancing function [1], a process called neofunctionalization. If the duplicated gene already had multiple roles (is pleiotropic), their partitioning among the duplicates would make each essential, a model called subfunctionalization [4]. Appeals to variations amongst duplicate genes can, in diploid species, also be applied to allelic variations in singleton genes [5]. The loss of one or both allelic function has been the basis of debates on whether dominance in Mendelian inheritance is an evolutionary or physiological phenomenon [6]. Dominance and gene duplication have both been framed [7] as phenomena that involve gene dosage – the contribution of the number of functional genes to phenotype – and the functional redundancy and fitness of genes may involve quantitative factors. Quantitative considerations include the disruption of stoichiometrically balanced protein levels when gene duplication increases the expression of one interacting partner [8]. It is the network of interactions that mediate the causal pathways from genes to phenotype and consequent evolutionary outcomes.

Novelties in evolution often emerge via changes to an organism’s development. A common mechanism in developmental trajectories is the transformation of transient stimuli into steady-state expression levels [9; 10; 11; 12] by gene regulatory circuits that implement positive feedback, wherein a gene upregulates its own expression. Such an autoregulatory gene activator formed the basis of an experimental study [13] on the “reversal of subfunctionalization” in the pathway that governs brain-stem development. While different developmental triggers activate the paralogous *hoxa1*, *hoxb1* gene pair in modern mice, they were replaced by a single autoregulatory activator responsive to a common set of *cis*-regulatory sites *cis*-regulatory inputs, and the resulting organism was viable [13]. It is this circuit of a self-regulating activator, the smallest unit of positive feedback, that is the

object of our study.

We study the consequences of duplication of this gene, both at the level of phenotype and in potential evolutionary outcomes, upon mutation. When mutations appear in this circuit as per the subfunctionalization model, introducing complementary loss of function *cis*-regulatory mutations that confers tissue-specific expression patterns [4; 14], we find the emergence of a property of some developmental trajectories, namely redundancy of duplicates helps overcome mutational loss [15]. The mutation of a gene expressed in a tissue is compensated for by the upregulation of its unexpressed partner in this model, called transcriptional backup [16]. Mutations affecting the activating sites of the transcription factor give rise to the onset of oscillations, both stable and bursty. Since the mathematical model for oscillations is also applicable to a diploid genotype with two alleles differing at their activating site, we examine the evolutionary fate of the duplicated gene haplotype in a diploid species via the fitness [17; 18; 7] of the oscillatory phenotype. Fitness values are indexed by parameter values in the doubled activator circuit that give rise to distinct qualitative dynamics.

The dynamical system we study is that of a diploid model of the transcriptional network [19], which enables us to consider both the interactions between diverging duplicates as well as allelic interactions [20], particularly those between allelic variants or heterozygotes. Heterozygous advantage has been identified as a property that facilitates the fixation of a duplicate gene in a population [21; 17; 5], a finding of relevance to our results below. Heterozygosity is also commonly associated with hybrid vigour; indeed, this correlation has also been extended to dosage (im-)balance of copy number variations and alterations to the amplitude of circadian rhythms [22]. Unlike circadian clock models [23], which contain a negative feedback link [24] in circuit topology, our duplicated activator network does not introduce negative feedback explicitly. Instead, negative feedback arises due to a competitive mechanism whereby one activator excludes the binding of another to the promoter. Although the mechanism implementing negative feedback – and thus oscillations – is different, our model displays a dependence of amplitude on heterozygosity and dosage balance found to be correlated to hybrid vigour [22]. Furthermore, the presence of paralogous genes in oscillatory processes has been noted in the literature [25; 26], also as a means to maintain oscillations upon loss of single genes [27; 28]. Although our model makes no explicit reference to the systems that contain the genes reported there, our theoretical finding suggests an evolutionary mechanism for the proliferation of duplicate genes that take part in oscillatory dynamics as heterozygotes.

There are two stages of modelling that we perform in this paper. The

key first step is to frame the onset of qualitative changes to the dynamics of transcription regulation as a consequence of gene duplication in the language of dynamical systems. The second step is to model the likely evolutionary fate of a mutant haplotype containing such a duplicate. For the first step, we set up a model of transcriptional regulation of a genotype characterised by a duplicated positive feedback loop, with the transcription rates derived from the probability of promoter occupancy by RNA polymerase [29] in the presence of the activators. An extension of the model to include the effects of intrinsic noise – stochasticity in transcription – is presented, to address how single cell and population averaged phenotype might differ. To enable the second stage of analysis on the evolutionary fates, a population genetic model for the likely invasion of a duplicated gene is then introduced, with selection coefficients that depend on parameters of the transcriptional network. Thereafter, we present results on the behaviour of these models. We identify qualitative shifts in the dynamics owing to alterations in parameter values – in *cis* and in *trans* – to be presented separately. After a discussion of the results owing to *cis*-regulatory changes that affect switching behaviour, we address the case of changes that affect regulation by *trans*-acting effects. It is this set of changes that we shall track in the second stage of modelling, that of the evolutionary fate of the duplicated gene. A final section summarises the different components discussed within the regulatory model, linking the qualitative aspects of model behaviour to different experimental studies.

## 2. Model of the duplicated autoregulatory gene switch

A positive feedback loop provides a mechanism to convert transient input signals into stable output levels, acting as a switch. Developmental stages, characterised by stable expression levels of subsets of genes, rely on regulatory circuits that implement positive feedback switches [9; 10; 11]. The smallest circuit implementing positive feedback is one with a single autoregulatory gene (see Figure 1); for it to act as a switch, it is necessary for the activation reaction of a transcription factor binding to the gene promoter to have a cooperativity index, or Hill coefficient, of 2 or more [30]. As a specific instantiation of such cooperativity, we require that the autoregulatory gene at the top of the hierarchy activates itself after dimerisation of its protein product (as shown in Figure 1), giving rise to a Hill coefficient of 2, although nothing in the model requires such a specific reaction. In particular, we disregard consideration of heterodimeric or homodimeric association, and assume the existence of homodimers alone. As will be clear

from the analysis below, a greater degree of cooperativity facilitates many more qualitative changes to the dynamics, which we shall disregard in the interest of parsimony. Our model takes such an autoregulatory gene coupled to target genes, which makes up a topology of the “terminal selector” network type [11], and duplicates it (see Figure 2). When duplicating the gene, we shall assume that both coding and regulatory regions of the gene are duplicated; thus, we end up with the network on the right hand side of Figure 2. Other influences on the activator might define developmental context or tissue specificity. Such interactions operationally outline the loci of context dependent changes that we shall introduce, and are indicated by the arrows labelled  $h$  into the activators  $a_{1,2}$  in Figure 2.

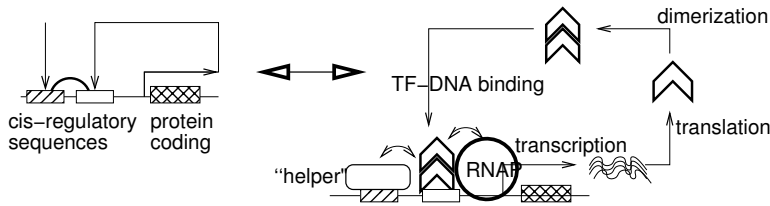


Figure 1: The haplotype on the left supports a representation of the schematic kinetics of transcription via dimeric activators which, with the adhesive reactions facilitated by the “helper” proteins, bind to DNA and recruiting the RNA polymerases. The mRNA transcribed is then translated, and homodimers are formed before the autoregulatory reactions proceed. There are further reactions that involve decay of mRNA and proteins that are considered in the models below.

### 2.1. Thermodynamic model of gene activation

The “thermodynamic” approach to modelling transcription is to set the rate of transcript formation to be proportional to the occupancy of the promoter of the gene to be expressed from a single allelic locus [29]. Promoter occupancy is facilitated by transcription factors that bind to cognate DNA sequences and recruit the transcription machinery – RNA Pol II, mediator complex forming multi-component proteins, *etc.* The probability of occupancy is accounted for by assigning Boltzmann factors for the possible configurations of protein-DNA bound states [29; 31]. The binding and unbinding protein-DNA reactions are assumed to be in rapid detailed balance to justify the use of the thermodynamic formalism; hence the ratios of these reaction probabilities are given by the negative exponential of the difference of free energies of the bound/unbound configurations.

The contribution of these configurations to the promoter occupancy is modelled in detail in Appendix A. Here we provide the resulting expression

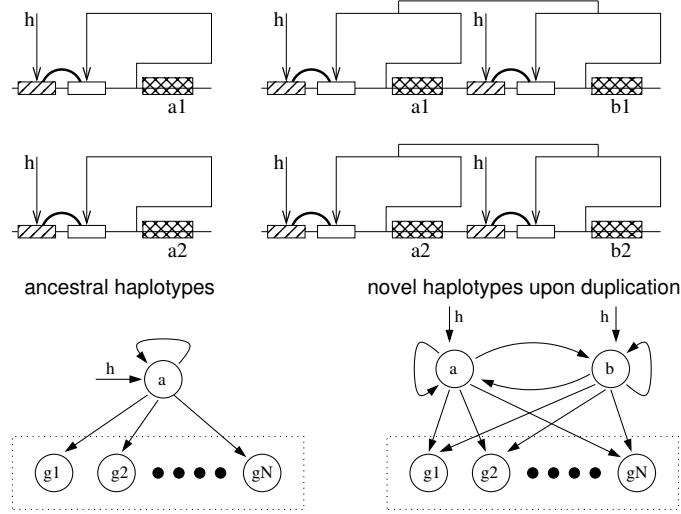


Figure 2: **Haplotype and network view of the duplicated gene.** The autoregulatory gene is part of the ancestral genotype and displayed in its network context on the left, with the allelic variants on top. The duplicated haplotype leads to an increased number of regulatory interactions in the network on the right.

for the probability  $\Phi_i$  of occupancy of the promoter of a gene  $i$ . This can be introduced in terms of the function  $\Psi_i$  of the transcription factors  $\alpha_1, \alpha_2$  (represented in vector notation  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ )

$$\Psi_i(\boldsymbol{\alpha}, \mathbf{r}_i, \mathbf{t}_i) = \frac{r_{i0} + t_{i1}r_{i1}\alpha_1 + t_{i2}r_{i2}\alpha_2}{1 + t_{i1}\alpha_1 + t_{i2}\alpha_2} \quad (1)$$

which measures the amount of transcript produced in the presence of the  $\alpha_i$  relative to the basal rate of transcription  $r_{i0}$ , when transcription factors are absent, a quantity called fold-change. The parameters  $r_{i1}, r_{i2}$  in  $\mathbf{r}_i = (r_{i0}, r_{i1}, r_{i2})$  stand for the strength of recruitment of the transcription machinery (Pol II, Mediator, etc) [32] by  $\alpha_1$  (rate  $r_{i1}$ ) and  $\alpha_2$  (rate  $r_{i2}$ ) from the chromosomal locus indexed by  $i$ . The parameters  $t_{ij}$  in  $\mathbf{t}_i = (t_{i1}, t_{i2})$  stand for *cis*-regulatory binding strengths of protein ( $\alpha_j$ ) binding to DNA locus ( $i$ ), determined by a more detailed description in eq. (3) below. The probability of promoter occupancy is given by

$$\Phi_i(\boldsymbol{\alpha}, \mathbf{r}_i, \mathbf{t}_i) = \frac{1}{1 + \Psi_i^{-1}(\boldsymbol{\alpha}, \mathbf{t}_i, \mathbf{r}_i)} \quad (i = 1, 2) . \quad (2)$$

The derivation of these expressions for  $\Phi_i$  and  $\Psi_i$  follows [33; 31], and is

presented in Appendix A. There we also derive the expression 3 below, which includes a “helper” protein that serves as a proxy for contextual influences. While the  $A_j$  indicates the proteins of interest (the duplicated activators) the “helper” proteins  $h$  are introduced to indicate the presence of co-regulators that enable tissue or developmental context-specific expression. In the following expressions (3), various  $\varepsilon$ s denote binding energies:  $\varepsilon_{Ad}^0$  denotes the non-specific binding of protein  $A$  to any of  $N_{ns}$  DNA binding sites;  $\varepsilon_{A_k d_k}^s$  stands for the binding of  $A_k$  to its cognate site  $d_k$ ;  $\varepsilon_{AB}$  measures the energy of protein-protein interactions between  $A$  and  $B$ ;  $\varepsilon_{hA_j p}$  the glue-like interaction between helper  $h$ , activator  $A_j$  and the RNA polymerase/Mediator complex  $p$ :

$$\begin{aligned}
\alpha_j &= \frac{A_j}{N_{ns}}, \\
r_{ij} &= \exp(-\beta \varepsilon_{A_j p}) \left( \frac{1 + \frac{h}{N_{ns}} \exp(-\beta(\varepsilon_{hd_i}^s - \varepsilon_{hd}^0 + \varepsilon_{hA_j p} - \varepsilon_{A_j p}))}{1 + \frac{h}{N_{ns}} \exp(-\beta(\varepsilon_{hd_i}^s - \varepsilon_{hd}^0))} \right) \\
t_{ij} &= \exp(-\beta(\varepsilon_{A_j d_i}^s - \varepsilon_{A_i d_i}^s)) \left( 1 + \frac{h}{N_{ns}} \exp(-\beta(\varepsilon_{hd_i}^s - \varepsilon_{hd}^0)) \right).
\end{aligned} \tag{3}$$

The details of the derivation is provided in Appendix A.2.

Eq. (3) parameterises are the molecular interactions that determine transcription rates in the gene regulatory network. While most of the paper will treat the  $r_{ij}$ ,  $t_{ij}$  (and other rates to be introduced shortly) as the parameters that determine network behaviour, the explicit definitions (3) reveal the substrates upon which mutations may act in order to change the protein-DNA and protein-protein interactions that determine phenotypic outcomes and evolutionary fates. In particular, in order to present verbal arguments that rely on mutations that disrupt pleiotropic properties and restrict them to context-specific roles (as in the subfunctionalization model of [4]; see also [34]) the helper protein index  $h$  will serve as a placeholder for such context-specific factors. This variable will be inherited by both copies of the gene upon duplication, and contribute to the set of mutations in *cis* (labelled by  $i$ ) by altering binding energy terms carrying both  $i$  and  $h$  indices. Having multiple helpers allow *cis*-context dependent effects by altering their binding affinities to DNA, and in particular in complementary ways – one helper (and not the other) binds to locus 1 and not to locus 2 and vice versa. We shall also study the effects of mutations that affect sites recruiting the transcriptional machinery and bring about changes in *trans*-acting terms  $\varepsilon_{A_j p}$  and/or  $\varepsilon_{hA_j p}$ , with the case of multiple helper proteins detailed in Appendix

A.4. Measuring binding energies relative to  $\beta = k_B T \sim 0.6$  kcal/mol, one can estimate the scale of the changes to parameters  $(t_{ij}, r_{ij}) \rightarrow (t'_{ij}, r'_{ij})$ . In particular, a change in binding free energy of  $\Delta\Delta G = 1.5$  and 2.5 kcal/mol, typical of biophysical measurements for mutant effects, translates to ratios of  $t_{ij}/t'_{ij}$  or  $r_{ij}/r'_{ij}$  of  $\sim 12$  or 65 respectively. These are the parameter ranges for which qualitative changes to transcriptional dynamics are noted below. The actual parameters used to perform the computational experiments as in Appendix B.

## 2.2. Model of haploid transcriptional deterministic kinetics

In the thermodynamic model for the rate of transcript production, we make the assumption that the transcriptional activators  $A_i$  act as homodimers, as shown in the representation of the autoregulatory circuit in Figure 1 that we shall take to be the ancestral haplotype. Dimer formation by monomer binding and unbinding is assumed to rapidly be in detailed balance on the time scale of gene expression, so that the concentrations of dimers  $[A_i] = x_i^2/K_{dim,i}$  where  $x_i$  is the concentration of monomers and  $K_{dim,i}$  is the dissociation constant for dimerisation. Monomers are translated from transcripts which decay at rates  $\delta_i$  much greater than the decay rates  $\Delta_i$  of the corresponding proteins. This assumption of faster time-scale of mRNA dynamics leads to the consequence that mRNA levels are slaved to protein dynamics. The equations for the rates of changes to proteins  $x_1$  and  $x_2$  thus captures the essential dynamics of the system. Here we present the model for the haploid case, with the detailed derivation provided in Appendix A.3:

$$\begin{aligned} \dot{x}_i &= c_i \varphi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i) - \Delta_i x_i, \quad (i = 1, 2), \\ \dot{\mathbf{y}} &= \mathbf{f}(x_1, x_2, \mathbf{y}) \end{aligned} \quad (4)$$

where  $\mathbf{f}(x, \mathbf{y})$  is the dynamical sub-system for the set of downstream variables  $\mathbf{y}$  that the activator gene  $x$  influences,  $\varphi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i) = \Phi_i(\mathbf{x}^2, \mathbf{r}_i, \mathbf{t}_i)$  in eq. (2), with  $\Delta_i$  the linear degradation rates for the proteins. The variables  $x_i$  have been scaled in terms of protein-DNA and dimer dissociation constants (see Appendix A.3), and the parameters  $c_i$ ,

$$c_i = \frac{[\phi_i] \pi_i \mu_i}{\delta_i} \quad (5)$$

are defined in terms of the rates for translation ( $\pi_i$ ) and mRNA degradation ( $\delta_i$ ). The transcription rate  $\mu_i$  expresses the proportionality between promoter occupancy in the thermodynamic description at each chromosomal locus, and  $[\phi_i]$  captures the number of such loci, which will play an important role in our later discussion on haploid and diploid cases.



Table 1: Reaction scheme for stochastic expression of proteins  $A_i$ ,  $i = 1, 2$ . There are 3 ways in which they are produced, by basal transcription, by regulated recruitment of transcriptional machinery by  $A_1$  or by  $A_2$  followed by translation. In the above,  $\Gamma_i = (1 + r_{i0}) + (1 + r_{i1})t_{i1}(\frac{A_1}{\kappa_1\Omega})^2 + (1 + r_{i2})t_{i2}(\frac{A_2}{\kappa_2\Omega})^2$ ,  $\Omega$  is the volume factor and  $\kappa_i$  is the geometric mean of the dissociation constants for dimerisation ( $K_{dim}^i$ ) of protein  $A_i$  and for protein-DNA binding  $\kappa_i^2 = K_{dim}^i \exp(\beta(\varepsilon_{A_i d_i}^s - \varepsilon_{A_i d}^0))$ .

reactions	rate of reaction
protein production:	
$\phi_i \xrightarrow{\text{basal}} A_i$	$c_i \frac{r_{i0}}{\Gamma_i \Omega}$
$\phi_i \xrightarrow{A_j} A_i$	$c_i \frac{r_{ij}}{t_{ij} \Gamma_i \Omega} \left( \frac{A_j}{\kappa_j \Omega} \right)^2$
protein degradation:	
$A_i \longrightarrow \emptyset$	$\Delta_i$

### 2.3. Stochastic kinetic model

The development of novel experimental techniques for tracking gene expression in single cells has made opened up for observation the consequences of the stochastic nature of the dynamics of gene regulation on phenotypes and their evolution[35; 36; 37]. In this section we present a simplified model of gene expression in this doubled autoregulatory gene network. This will enable us to explore the consequences of intrinsic regulatory noise on the phenotype and what its implications might be for evolutionary fates of the duplicate genes. While the detailed model of stochastic kinetics is provided in Appendix A.3, here we present a simpler model that incorporates many of the reaction steps into a Hill-type gene regulatory function, just as in the deterministic version in eq. (4). The reactions are summarised in Table 1.

### 2.4. Diploid model of regulation

In diploid organisms, each gene comes in two copies independent of any duplication event. To examine how duplication generates novelty, or to track how a genotype containing a mutant duplicate can be subject to evolutionary modelling, we need to construct a model of transcriptional regulation that can track allelic variants. Further, we restrict differences between alleles to be solely in their activation sites, with different alleles (both before and after duplication) affecting transcription rates only by differing affinities for the transcriptional machinery –  $\varepsilon_{A_1 p} \neq \varepsilon_{A_2 p}$  and  $\varepsilon_{hA_1 p} \neq \varepsilon_{hA_2 p}$ . Consequently the coefficients in  $\mathbf{r}_i = (r_{i0}, r_{i1}, r_{i2})$  are independent of  $i$ . To restrict our

attention to activation site changes, based on results obtained from model analysis below, we shall assume that there are no differences between the alleles in the free energies of protein-DNA binding, hence we take  $\mathbf{t}_i = \mathbf{t}_j = 1$ , where we have rescaled the concentration variable to be multiples of the protein-DNA and dimer dissociation constants that make up  $\mathbf{t}_i$  (see Appendix A.3). While the dynamics of transcriptional regulation will be studied in both  $i$ -dependent and  $i$ -independent cases below (where  $i$  is the genomic locus), the evolutionary dynamics of duplicates will be explored only in the context of mutational variations affecting activating sites in a modular fashion, independent of *cis*-regulatory context.

#### 2.4.1. Singleton case

If there are two allelic forms  $x_i^{(j)}$  where  $j = 1, 2$  for each gene  $x_i$  we can extend the development leading to (4) to consider how each allele generates a transcript at the rate determined by its promoter occupancy by transcription factors. We can now set up a model to investigate variability in transcription dynamics owing to allelic variations, or heterozygosity for the autoregulatory circuit prior to duplication. We introduce an allelic index  $j = 1, 2$  in the fold-change  $\Psi_i^{(j)}$  for each gene labelled by  $i$  [19]. For the case of a single locus  $i = 1$  with the corresponding fold change  $\Psi_1^{(j)}(x_1^{(1)}, x_1^{(2)})$ . If the two alleles  $x_1^{(1)}, x_1^{(2)}$  are the same, the homozygous case, the diploid genotype network is shown in Figure 3 (A), but the model dynamical model is equivalent to the ancestral haplotype in Figure 2(a), or the switch model of a single autoregulatory gene, but with the parameter representing protein-DNA binding doubled:  $t \mapsto 2t$  (eq. (1)). For the heterozygous case, allelic variants  $x_1^{(1)}, x_1^{(2)}$  may be regarded as the two distinct activators  $A_1$  and  $A_2$  in the haplotype depicted in Figure 2(b). The dynamical equations are then the same as in the haploid case, except that  $x_1^{(1)} \mapsto x_1$  and  $x_1^{(2)} \mapsto x_2$  in eq. (4).

#### 2.4.2. Duplicated allele

There are two cases to consider when one of the alleles is duplicated – if the ancestral haplotype is homozygous or heterozygous. If homozygous, the duplicated genotype is again of the original haplotype topology as in Figure 2(a). There is thus only one species of transcription factor protein which regulates itself and its downstream components  $y$ , with the effective dissociation constant to its DNA binding site scaled by a factor of 4 relative to the single copy haploid model. This will only shift the threshold for switching in the analysis presented below.

If the ancestral genotype is heterozygous, containing two different alleles  $a_1$  and  $a_2$  at a locus labelled by  $A$ , duplication creates another chromosomal locus,  $B$ , in a mutant haplotype containing 2 copies of one genetic sequence, say  $a_1b_1$ . The corresponding genotypes can in a mixed mating population, considered in greater detail below, are of the form  $a_i/a_j$  (4 combinations for  $i, j = 1, 2$ )  $a_ib_j/a_k$  (8 combinations) and  $a_ib_j/a_kb_l$  (16 combinations). All these combinations require only 2 different species of transcription factors. If the allelic variant is in a coding region, the differences show up in the rates of recruitment of the transcription machinery,  $\varepsilon_{A_jp}$ , whereas if the mutations are in the regulatory region, the differences to be considered are in the binding of helper/coactivator proteins, that change  $\varepsilon_{hA_jp}$  in eq. (3). (The inability to form heterodimers between distinguished alleles is still assumed.) Upon taking these changes into account, the difference in the number of distinct gene copies to be considered is contained in the parameter  $[\phi_i^{(k)}]$  that provides allelic specificity to  $c_i$  in eq. (5).

In sum, the corresponding model is described by the ODEs

$$\begin{aligned} \dot{x}_i &= \sum_k c_i^{(k)} \varphi_i^{(k)}(\mathbf{x}, \mathbf{r}_i^{(k)}, \mathbf{t}_i^{(k)}) - \Delta_i x_i, \quad (i, k = 1, 2), \\ &= \sum_k c_i^{(k)} \varphi_i^{(k)}(\mathbf{x}, \mathbf{r}^{(k)}, \mathbf{t}^{(k)}) - \Delta_i x_i, \end{aligned} \quad (6)$$

where (again) the superscript  $(k)$  refers to the two alleles and their corresponding parameters refer to possibly different binding interactions at each allele as introduced. The parameters  $c_i$  in (eq. (5)) can be extended to reflect potential allelic differences in transcription, translation or degradation rates for proteins and transcripts, but which we keep the same for both copies of the duplicated allele, since our interest lies in the mutation that leads to an allele with two copies. Thus, to present the results in the latter sections, it is the model in eq. (4) that will be studied with the parameter  $c_i$  playing a role in the evolutionary discussion.

### 2.5. Adaptive dynamics of duplicated gene

Most of what we have modelled of the autoregulatory switch has been independent of the mode of its inheritance, to which we have appealed to in order to set up the terms of discussion, but which we have not presented in detail to pursue its possible evolutionary implications. In particular, our discussion has implied a continuity of argument from haplotype to inheritance and evolutionary fate. As noted, the key difference introduced by duplication at the level of haplotype is the multiple feedback structure, an interaction topology that is novel for haploids but not diploids. For the diploid case, the locus of change for system dynamics upon duplication is

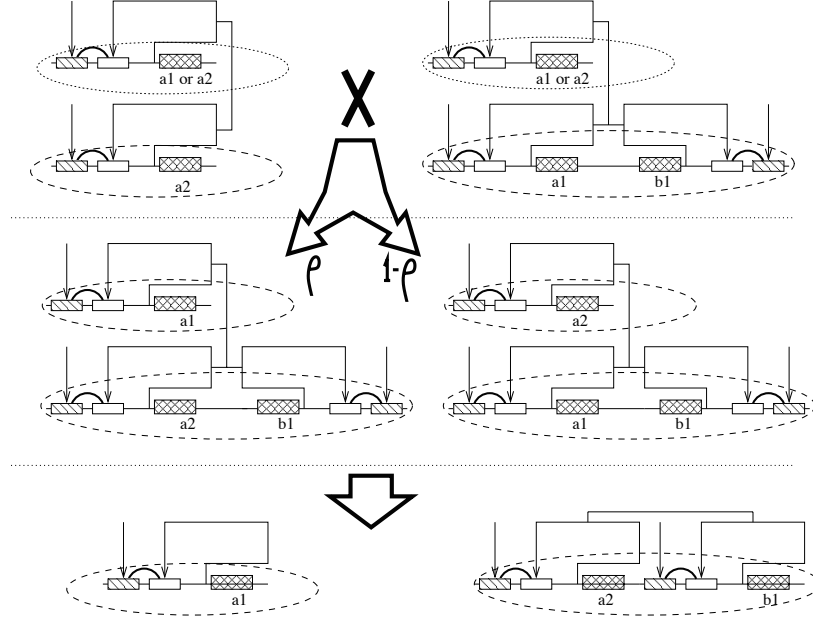


Figure 3: **Recombination amongst haplotypes and topologies for diploid regulatory networks.** Mating between two diploid genotype networks, one containing a chromosome with a duplicated segment of DNA, is shown with an X. The diploid network with duplicate genes on both chromosomes is not shown as that occur with a very small probability immediately after the duplication event. A duplicate segment of DNA gets shuffled by the process of recombination, which is assumed to occur at a rate  $\rho$ . Of the mating genotypes, the alleles chosen in the gametes are the dotted and dashed pairs, of which the dashed pair is tracked in the figure as it undergoes recombination events. In the model presented below, singleton alleles are at their equilibrium frequencies in the population.

tracked by the parameter  $c_i$  which incorporates the copy number  $[\phi_i]$ . The number of functional copies of the gene used for Mendelian arguments and for matching their effects to the environments inhabited by the phenotype ranges from 0 to 4 in the diploid case [7]. Recombination in sexually reproducing diploid organisms “randomises” novel mutations along genomes and may offer greater possibilities for fixation by selective forces [38]. In the previous section we have associated the analysis of divergent properties of post-fixation duplicates to pre-duplication heterozygotes [21; 5] by mapping them onto a common dynamical system for transcriptional dynamics. In this section, we investigate the evolutionary consequences of network level effects of homozygous and heterozygous genotypes using a model for the

evolution of duplication based on [17]. In this model, a mutant duplicated allele is introduced into, and examined for its ability to invade, a population characterised by an existing equilibrium gene pool. Equilibrium configurations provide standard population genetic backgrounds in which a mutant's ability to survive is analysed. The transcriptional network of interest in this paper makes it natural to consider the effects of distinct alleles in the equilibrium population; hence, we examine the case of equilibrium maintained by overdominance, where heterozygotes are fitter than either homozygote in this bi-allelic setting. We also do not wish to consider the effects of mutation after the duplication event in the evolutionary picture (as it is already a lengthy paper); thus, we eliminate consideration of equilibria involving mutation-selection and mutation-drift balance [39].

The model includes haplotypes defined via a pair of loci  $A, B$ , with allelic values of  $a_1, a_2$  for  $A$  and  $b_0, b_1, b_2$  for  $B$ .  $b_0$  is the null allele, so the haplotypes  $a_1b_0$  and  $a_2b_0$  are the alleles  $a_1, a_2$  prior to duplication. (Note, that we have introduced the null allele at the second locus for convenience. If we allowed for a null allele,  $a_0$  at the first locus, the pre-duplication alleles would, equivalently, be expressed as  $a_0b_1$  and  $a_0b_2$ .) In the overdominant case, the model assumes an initial polymorphic equilibrium configuration of alleles  $a_1$  and  $a_2$ , via heterozygote advantage [21]. These singleton haplotypes are also called  $a_1b_0$  and  $a_2b_0$  below. Homozygotes  $a_1/a_1$  and  $a_2/a_2$  (also named  $a_1b_0/a_1b_0$  and  $a_2b_0/a_2b_0$ ) have fitness coefficients  $1 - s$  and  $1 - t$  relative to the heterozygote  $a_1/a_2$  with a fitness coefficient of 1; consequently, the relative frequencies of alleles  $a_1$  and  $a_2$  at equilibrium are  $\hat{x}_{10} = t/(t + s)$  and  $\hat{x}_{20} = s/(t + s)$ , respectively [39]. The fitness of the population at this polymorphic equilibrium is  $\hat{W} = 1 - ts/(s + t)$ .

A duplicate mutant is introduced into a background of this existing equilibrium condition. The duplicate haplotypes are denoted  $a_1b_1, a_2b_1, a_1b_2, a_2b_2$  with frequencies  $x_{11}, x_{21}, x_{12}, x_{22}$  respectively. Following [17], we assume a two step decomposition for updating the haplotype frequencies:

$$x_{ij} \mapsto x_{ij}^* \text{ by recombination, and we shall ignore } x_{ij}^* \mapsto x'_{ij} \text{ by mutations.}$$

We set up the discrete dynamics of frequency updates due to recombination by first assuming that the mutant duplicate has a probability of being chosen for mating with a probability  $\sim N_e^{-1}$  where  $N_e$  is the effective population size, much smaller than those of the singletons  $\hat{x}_{10} = t/(t + s)$  and  $\hat{x}_{20} = s/(t + s)$  which are at their equilibrium frequencies. Therefore the probability of pairing duplicate gametes with each other is negligible compared with those of pairing a duplicate haplotype with that of a singleton

[17]. After duplicating  $a_1$  to produce  $a_1b_1$ , recombination, mating individuals containing gametes of haplotype  $a_1b_1$  with those containing  $a_2b_0$  occurs with probability proportional to  $\hat{x}_{20}x_{11}$ , creating a  $a_1b_0/a_2b_1$  genotype at a rate proportional to recombination probability  $\rho$  and  $a_2b_0/a_1b_1$  with probability  $(1 - \rho)$ . The fitness values of the two genotypes are indexed by their subscripts –  $W_{10;21}$  and  $W_{20;11}$  respectively. To first order, the equilibrium frequencies  $\hat{x}_{10}$  and  $\hat{x}_{20}$  are unchanged.

The discrete map is represented via the block-diagonal matrix:

$$\begin{pmatrix} x_{11}^* \\ x_{21}^* \\ x_{22}^* \\ x_{21}^* \end{pmatrix} = \frac{1}{\hat{W}} \begin{pmatrix} \mathbf{R}_{11 \times 21} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{22 \times 12} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{22} \\ x_{12} \end{pmatrix}, \quad (7)$$

where the  $(2 \times 2)$  submatrices are defined in (8) below.

$$\begin{aligned} \mathbf{R}_{11 \times 21} &= \begin{pmatrix} W_{10;11}\hat{x}_{10} + (1 - \rho)W_{20;11}\hat{x}_{20} & \rho W_{10;21}\hat{x}_{10} \\ \rho W_{20;11}\hat{x}_{20} & W_{20;21}\hat{x}_{20} + (1 - \rho)W_{10;21}\hat{x}_{10} \end{pmatrix}, \\ \mathbf{R}_{22 \times 21} &= \begin{pmatrix} W_{20;22}\hat{x}_{20} + (1 - \rho)W_{10;22}\hat{x}_{10} & \rho W_{20;12}\hat{x}_{20} \\ \rho W_{10;22}\hat{x}_{10} & W_{10;12}\hat{x}_{10} + (1 - \rho)W_{20;12}\hat{x}_{20} \end{pmatrix} \end{aligned} \quad (8)$$

The fitness allocation for  $a_i b_0/a_i b_0$  is taken to be the same as that of the  $a_i b_i/a_i b_0$  genotype, for  $i = 1, 2$ , so  $W_{10;10} = W_{10;11} = 1 - s$  and  $W_{20;20} = W_{20;22} = 1 - t$ . This assumes that, since identical genes can only generate switch-like behaviours, the outcome of increasing dosage is a shift in the threshold for switch activation only, which we assume to be neutral (see below). For the case where the copy numbers for the allelic variants of the single genes are in the ratio 2 : 1 or 1 : 2, we take  $W_{10;12} = W_{10;21} = W_{20;11} = 1 + d$  and  $W_{10;22} = W_{20;12} = W_{20;21} = 1 - u$ . We shall make a note of the case  $d = 0$  since we would like to impose fewer constraints for positive selection explicitly. All of these fitness coefficients are normalised with respect to the pre-duplication heterozygote,  $W_{10;20} = 1$ . The fitness coefficients are summarised in Table 2.

Following standard practice, we have summarised the effects of viability and reproductive success by single scalar-valued parameters. We will need to relate these parameters to dynamical states in the model of the autoregulatory gene activator in order to make claims about evolutionary consequences of regulatory changes. Thus,  $s, t, d, u$  will be defined in terms of the parameters in the reaction system shown in A.3 in the Appendix, and which for the purpose of dynamical analysis we have summarised in terms

Table 2: The fitness values  $W_{ij;kl}$  for the different genotypes obtained by mating  $ij$  with  $kl$  gametes. The selection coefficients are with reference to the singleton heterozygote which is assigned a fitness of 1. All coefficients  $s, t, u$  are positive, indicating reduced fitness compared to the  $a_1/a_2$  heterozygous singleton, while both signs of  $d$  are briefly explored in the main text.

$W_{ij;kl}$	10	20	11	21	22	12
10	$1 - s$	1	$1 - s$	$1 + d$	$1 - u$	$1 + d$
20	1	$1 - t$	$1 + d$	$1 - u$	$1 - t$	$1 - u$

of the parameter combinations that arise in (4). Hence the selection coefficients of genotypes  $a_k b_l$  capture dependence on the network parameters  $s = s(\mathbf{r}_i, \mathbf{t}_i, c_i, \Delta_i)$  in a specific environment via their dynamical behaviour. Since we do not model any particular environment in this paper, we will require verbal arguments indicating the plausibility of adaptive roles of the dynamical states in parameter space. The passage from a continuous set of biochemical parameters to a discrete set of selection coefficients will be motivated via the appearance of distinguished qualitative regions in phase space (see Figure 11 below) that partition the phenotype space into a discrete set of qualitatively different dynamical behaviour.

### 3. The two component subsystem feeds novelty downstream

In this section we will argue that the autoregulatory component of the network in Figure 2 can be analysed in isolation of its downstream effects for the kinds of effects that we will focus on – the effects of dosage (im)balance introduced by gene duplication. This will be obtained by showing how the eigenvalues of the linearised dynamics can be factorized. We then show an immediate consequence of competition between the two copies of a gene for regulatory binding and activation.

#### 3.1. Restriction to the two-component subsystem

The Jacobian of the dynamical system in eq. (4) has the same structure as the adjacency matrix in Figure 2). Since the Jacobian determines the local dynamical behaviour around any state of the system, changes to its eigenvalues signal the onset of qualitative behavioural patterns. In particular, around each fixed point of the dynamical systems of the gene regulatory networks shown in Figure 2, the Jacobian of the network gets updated as

shown :

$$\left( \begin{array}{c|ccc} a_{11} & 0 & 0 & 0 \\ \hline g_{11} & & & \\ \vdots & & & \\ g_{N1} & & & \end{array} \middle| \mathcal{G}_{N \times N} \right) \longrightarrow \left( \begin{array}{cc|ccc} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{12} & a_{22} & 0 & 0 & 0 \\ \hline g_{11} & g_{12} & & & \\ \vdots & \vdots & & & \\ g_{N1} & g_{N2} & & & \end{array} \middle| \mathcal{G}'_{N \times N} \right). \quad (9)$$

In eq. (9), the  $(\mathcal{G}_{N \times N}) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$  submatrix of the Jacobian corresponds to the part of the network enclosed in a dotted box in Figure 2;  $g_{i1}$  ( $i = 1, \dots, N$ ) are the partial derivatives of the promoter occupancy functions of the downstream genes with respect to the variable representing the autoregulatory gene.  $a_{11}$  denotes the partial derivative of the net rate of expression of the autoregulatory gene with respect to its own expression state. After duplication (or in the biallelic case), there are two genes that influence the downstream sub-network, and auto- and cross-dependencies in the regulatory dynamics of the activators with Jacobian elements  $a_{ij}$  are introduced. The block structure of the Jacobian matrices implies that the determinants are of the form

$$a_{11} \times |\mathcal{G}| \text{ and } \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \times |\mathcal{G}'|. \quad (10)$$

and thus the eigenvalues that determine dynamical consequences factor into two pieces. This modular decomposition suggests the following strategy.

The argument invoking the selective advantage of increased gene dosage [7] looks to the effect of the doubling of steady state levels  $x^* \rightarrow (x_1^* + x_2^*)$  on the target gene dynamics  $\mathbf{f}(x^*, \mathbf{y})$  (where we use  $*$  to indicate steady-state levels in the regulatory network in a feed-forward manner. These changes  $\mathbf{f}(x_1^* + x_2^*, \mathbf{y}) - \mathbf{f}(x^*, \mathbf{y})$  may indeed be associated with a positive selection coefficient. However, we shall instead look at other sources of qualitative shifts – not gene dosage, but gene dosage balance instead. We shall assume that the autoregulatory positive feedback serves to set a binary decision switch for the downstream components to be triggered, and the net effect of this doubling ‘merely’ adjusts the threshold of the switch. Since we look for signals of dynamical shifts via the local analysis provided by the eigenvalues of the Jacobian, and the Jacobian in eq. (9) is of a factorised form, the evolutionary significance of the changes ascribed to downstream sensitivity to doubled gene dosage lies in  $|\mathcal{G}'|$  in eq. (10). In this paper, we are interested in the qualitative changes that might influence phenotypes of duplicated genes lies in the other factor involving the  $a_{ij}$  components. In considering dynamical changes, such as the onset of oscillations to be



considered below, this explains why we focus on the two-gene sub-network involving the duplicated autoregulator. For the case of the switch, we shall use the downstream levels as a reporter of the dynamical switches in the duplicated gene motif.

The set of reductions required to get to this simplified form (4) assumes separation of time-scales of transcript and protein formation and degradation relative to protein-DNA and protein-protein interactions. This can introduce differences in the time scales of the results, particularly when dimerisation is involved [40] – the full system has slower dynamics which shows up in its time-dependent behaviour. Similar consideration needs to be paid to the time scales of the interactions that couple the double-activator motif to the downstream components. It is known that coupling to downstream components can alter qualitative dynamics of a network, a principle dubbed “retroactivity” [41]. We have checked that the dynamics of the coupling can indeed affect the behaviour of the system, as does the kinetics of dimerisation. However, we have checked that it is possible to find parameter ranges for downstream coupling and dimerisation kinetics for which the behaviour of the full system behaves in a manner qualitatively similar to the simplified model we choose to focus on, albeit with a slower time-scale for the dynamics. However, for the purposes of this paper, these differences are not significant; all the essential qualitative features predicted from this simple model are also present in the full kinetic description.

For the analysis below, we shall focus on two principal cases for the combinatorial regulatory parameters  $\mathbf{t}_i, \mathbf{r}_i$ . In one, we set  $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$ , corresponding to the case when recruitment for activation is modular, *i.e.* independent of *cis*- context *i*. Thus,  $(r_{10}, r_{11}, r_{12}) = (r_{20}, r_{21}, r_{22}) = (r_0, r_1, r_2)$  and  $\varphi_1 = \varphi_2 = \varphi$ . When *cis*-regulatory context matters for expression, for instance, in the scenario described in the subfunctionalization model, we consider  $\mathbf{r}_1 = (r_0, r_1, r_2)$  and  $\mathbf{r}_2 = (r_0, r_2, r_1)$ . If  $r_1 > r_2$ , this choice indicates a greater rate of transcription of gene 1 from duplicate locus 1 and a greater transcription rate for gene 2 from locus 2. This can be achieved by assigning different affinities for the helper proteins *h* in the two contexts labelled by *i* in equation (3). A detailed derivation of how this emerges is provided in Appendix A.4, and Figure 4 makes the modelling assumptions explicit.

### 3.2. Dual regulation as a consequence of changes to activation domain

The fold-change function  $\Psi_\ell$  (2) determines whether the influence of a protein on gene expression is that of an activator or a repressor, indicated by

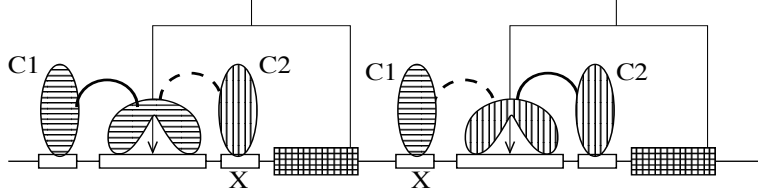


Figure 4: **Binding strengths and *cis*-regulatory mutation leads to the heterozygous switch.** The two proteins expressed by the two alleles/duplicate genes are shown in vertical and horizontal stripes bound to their DNA binding site indicated by the target of the regulatory arrow. They have complementary binding strengths to the helper proteins  $C_{1,2}$  appropriately shaded and further indicated by the arcs between the proteins – the dark arcs denote greater strength of association compares to the dashed ones. The sites marked X are where complementary loss of binding occurs upon mutation, as is postulated in the subfunctionalization model [14].

the sign of its derivative with respect to the amount of transcription factor. The derivatives of  $\Psi_\ell$  with respect to  $\alpha_{1,2}$  is given by

$$\left( \frac{\partial/\partial\alpha_1}{\partial/\partial\alpha_2} \right) \Psi_\ell(\alpha, \mathbf{t}_\ell, \mathbf{r}_\ell) = \frac{t_{\ell 1} t_{\ell 2} (r_{\ell 1} - r_{\ell 2})}{(1 + t_{\ell 1} \alpha_1 + t_{\ell 2} \alpha_2)^2} \left( \frac{\frac{r_{\ell 1} - r_{\ell 0}}{t_{\ell 2} (r_{\ell 1} - r_{\ell 2})} + \alpha_2}{\frac{r_{\ell 2} - r_{\ell 0}}{t_{\ell 1} (r_{\ell 1} - r_{\ell 2})} - \alpha_1} \right), \quad (11)$$

flagging the possibility of non-monotonicity of  $\Psi_\ell$ . This non-monotonicity means that increasing the amount of a transcription factor produces a fold change of transcription that increases (activates) in one context and decreases (represses) in another, a feature called dual regulation. In this model, the context is set by the concentrations of the paralogous protein. We consider the case  $r_{\ell 1} > r_{\ell 2}$ , where the  $A_1$  binds more strongly to the Pol II enzyme than  $A_2$  (ignoring enhancer context  $\ell$  for  $h = 0$  in (3)) and find that  $A_2$  behaves as an activator at low levels of  $A_1$  and a repressor when the  $A_1$  level  $\alpha_1$  crosses a threshold [42; 43]. This repressor-like behaviour of  $A_2$  (the activator with weaker affinity to Pol II) occurs even when  $A_2$  is a facilitator of transcription by itself. Since access to the binding site on the DNA is a limiting resource, with both activators competing for it, increasing the levels of the weaker activator hampers the overall efficiency of transcription from the combined  $(A_1, A_2)$  system. Unequal binding affinity to DNA target sites ( $t_{\ell 1}, t_{\ell 2} \neq 1$ ) changes the amount of regulator  $A_1$  that must be present for the crossover behaviour to occur. Such dual regulation can also be a property of activators expressed from a single polymorphic locus.

This feature, that duplication of an autoregulatory gene can introduce

competition for regulatory sites on the DNA and lead to dual regulation, opens up the possibility that novelty can arise due to mutations to modular components of one of the duplicated proteins which alters  $r_{ij} = \exp(-\beta\varepsilon_{A_{jp}}) =: r_j$  or  $\exp(-\beta\varepsilon_{hA_{jp}})$  or even the DNA binding energy of the helper protein as seen in eq. (3). In all of what we discuss in this paper, the key parameters that we change are the  $r_{ij}$  in eq. (3) – the arrow connecting transcription factor to the “helper” in Figure 1 – while we keep  $t_{ij}$  to be the same across the two genes of interest, just for simplicity. The  $r_{ij}$  do depend on protein-DNA binding strengths, but only for the helper proteins, as shown in Appendix A.4.

#### 4. Changes to dynamics of the network – switches and backup

The autoregulatory circuit behaves like a switch because the dynamical system (4) supports 2 stable fixed points for a given choice of parameters. Initial conditions specifying protein levels on either side of a threshold value drives the system to its low or high expression state. In this section we consider the cases where having two genes coupled via feedback gives rise to different types of switches, referred to in [44] as a progression switch (where two genes can move from low to high expression states) or a decision switch (where the genes switch to low-high or high-low expression states, the simplest example of a choice of regulatory fate).

##### 4.1. Homozygous or progressive switch: lowering the on-off threshold

Duplication of a gene has typically been associated with increase of protein product. If the gene were part of a switching circuit with a phenotype that is binary-valued, as is the case for networks that convert transient, threshold-crossing inputs into stable outputs, it might well be the case that doubling of a gene does not greatly increase protein product, but merely makes the threshold more accessible for crossing. We illustrate this possibility by setting  $r_{ij} = r$  and  $\Delta_i = \Delta = 1$  and  $c_i = c = 1$  in (4). This is equivalent to a single gene switch but with  $t_{ij}$  doubled. As expected, altering the dissociation constant for an autoregulatory circuit changes the threshold in a sigmoidal function. The steady states are solutions to a cubic equation with two stable fixed points separated by a saddle. The bifurcation from a mono-stable to a bistable state is of a saddle-node variety [45]. The nullclines, bifurcation plots and histograms of expressed proteins generated from a Gillespie simulation make the point vividly in Figure 5. This subsection is presented only to make the comparison with the asymmetric case obvious.

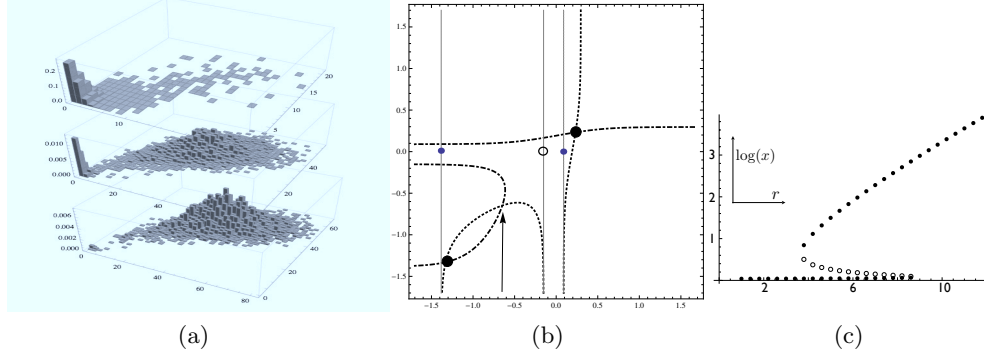


Figure 5: **The location of a transition from a low expression state to a high expression state is lowered upon duplication with no further divergence.** On the left are histograms of the expression levels of the duplicated self-activating gene, with the bifurcation from one to two to one stable steady state for changing  $r_{ij} = r$  from 1 through 10 to 20 (from top to bottom). All other parameter values are as in Appendix B, with  $c = \Delta = 1$ , and the activation strengths of the (identical) proteins taken as  $r \times r_0$ . The location of the modes (specifically shown for the histogram in the middle) correspond to the intersections of the (dotted, dot-dashed) nullclines indicated by the big black circles on the right. The vertical lines in (b) correspond to the fixed points for the single autoregulatory loop, with the smaller dots indicating steady state levels there. The open circle on the vertical line in the middle and the intersection indicated by the arrow are the locations of the thresholds in the single and duplicate gene cases respectively. In (c) the abscissa denotes increasing values of  $r$  while the ordinate is the expression level. The characteristic S-shape of a saddle node bifurcation with hysteresis is seen. The  $x_1$ - $x_2$  planes in this figure are in logarithmic scale.

#### 4.2. Heterozygous or decision switch via complementary loss of recruitment

In this case, we note that the two duplicate proteins recruit Pol II with the help of complementary helper proteins at the two loci. Gene copy 1 is activated by protein 1 at a rate that is greater than that achieved by duplicate protein 2,  $r_{11} = r \times r_{12}$  with  $r > 1$ . The roles are reversed for expression from gene copy 2, with  $r_{22} = r' \times r_{21}$  with  $r' > 1$ . The detailed origin of these parameters lies in the loss of complementary regulatory binding sites of two helper proteins that have complementary binding preference to proteins of the two alleles/genes, as explained in Appendix A.4. They are summarised in the Figure 4. The combined strength of recruitment of the transcriptional machinery is partitioned (as shown in Figure 4, and in detail in Appendix A.4) via loss of binding sites for the helper proteins. For simplicity, we shall take  $r_{21} = r_{12}$  and  $r = r'$ ; the symmetry does not introduce any non-generic features to this dynamics, but makes the analysis more transparent.

The fixed points are now given by intersections of cubics obtained by setting  $(d/dt)x_{1,2} = 0$  in eq. 4), where we ignore the downstream sector. We find multi-stable solutions in this case – a tri-stable state with  $(x_1^*, x_2^*)$  levels being (low, low), (low, high) and (high, low) which undergoes a change via a pitchfork bifurcation upon increasing  $r$  to a bi-stable state where the low expression state is lost and only the mutually exclusive expression states (low, high) and (high, low) for  $(x_1^*, x_2^*)$  remain. Figure 6 illustrates this.

We have thus far studied the effects of complementary mutations following duplication that affect  $r_{ij}$  by affecting protein-protein interactions at the enhancers and thus  $\varepsilon_{hA_{jp}}$  in eq. (3), where helper proteins indexed by  $h$  provide transcriptional response specificity in this modelling framework.  $h$  is a place-holder for the *cis*-regulatory context of gene expression. In order to implement the complementary loss-of-function mutations of [4], we have the helper protein enhance the recruitment of transcriptional machinery in complementary ways. Starting from  $\mathbf{r}_i = (r_{i0}, r_{i1}, r_{i2})$  with equal  $r_{i1}$  and  $r_{i2}$ , we end up with a situation where transcription from locus 1 is greater for the coding region 1 (say) than the coding sequence 2 ( $r_{11} > r_{12}$ ) and similarly,  $r_{22} > r_{21}$  (details in Appendix A.4). We compare the two cases of progressive (homozygous) and decision (heterozygous) switches [44] in Figure 7.

We use the expression level of a downstream gene as an indicator variable (abscissa in Figure 7) to locate the influence of the control parameter  $r$ , the ratio of activation strengths of the two transcription factors (ordinate in Figure 7). The filled circles correspond to stable fixed points of the expression of a downstream gene activated by the pair of duplicated activators and the open circles are those that correspond to the unstable fixed points identified in Figures 5, 6. In Figure 7 (c) we show that the key difference in the two cases (now the open circles correspond to the closed circles, or stable points in (b), and the filled circles are the stable steady states in the duplicated-no-divergence case of (a)) lies in the extent of hysteresis that the system affords. The doubled autoregulator shifts the threshold for entry into the high-expression state lower compared to the pre-duplicated gene. The asymmetrically expressed decision switch case is compared to the pre-duplicated situation in Figure 7 (d), with barely any difference in the expression levels in the two cases.

The doubling of the self-activating gene comes with an obvious corollary in that the system retains one copy of the switch upon deletion of one of the two copies of the gene. Thus, even in the case that a gene is in a low-expression state as in Figure 6 (b), the deletion of its high expression partner induces its upregulation. Such a back-up feature points to the plausibility

of invoking robustness in differentiating between the systems yielding the two expression levels in Figure 7 (d), which are otherwise indistinguishable. This stable dynamical state is of course, unstable from an evolutionary perspective unless the conditions that lead to the potential loss of a paralog are heritable as well.

An hypothesised original autoregulatory developmental gene has been used to replace two copies of the *hox1a,b* paralogs in a mouse with consequent minor alterations to normal development of its forebrain[13]. However, in the wild-type mouse, the prominent phenotypes that develop upon knock-out of a paralog means that this back-up facility that is a consequence of duplication has been lost, via the loss of the *hoxa1* autoregulatory site, and acquisition of further specialised roles for the proteins. This dynamically stable state of the duplicated autoregulator is rendered evolutionary stable by the divergence of the pleiotropic roles taken on by the duplicates, which are split and stabilised by the model of subfunctionalization [4; 46]. An example of a developmental system where genetic buffering upon deletion of a paralog is observed is the *myoD* and *myf5* pair involved in myogenesis where deletion of the pair induces the upregulation by its partner[15]. However, this appears not to be a cell-autonomous property: results from clonal lines *in vitro* do not show the same effects as in the organism, and further studies [47] indicate lineage-specific divergence of gene function, with buffering occurring at an inter-cellular stage.

## 5. Emergence of oscillations due to changes to activation site

The standard model of gene activation [32] involves recruitment of Pol II and other transcriptional proteins by protein-protein interactions with the activation domain of the transcriptional activator. Increasing the activity of the (acidic) activation domain is also accompanied by an increase in the protein degradation rate [48], a correlate that is key to the results below. Mutations to the activation domain to either copy of the duplicated gene/allele can thus introduce a change in the time scales of the dynamics of the two proteins. If we view the enhancer region of the gene as the source from which proteins are produced, duplication followed by alteration of the activation region introduces competition for this source. Such a competitive framework arises in ecological theory where one species can take over a food source or habitat at the expense of another – a property called competitive exclusion. However, it has been shown [49] that it is possible for two competing species of predators to subsist on a single species of prey (a food source) in an oscillatory mode. This analogy extends to our model as well.

For the case where the divergence between copies is modular both enhancers are taken to be copies of each other with  $t_{ij} = 1$  and the occupancy of the promoters are the same,  $\varphi_1 = \varphi_2 = \varphi$ . In detail, modular activation implies  $r_{ij} = r_{oj}$ , independent of genomic locus  $i$ . Thus, we take

$$\mathbf{r}_i = (r_{i0}, r_{i1}, r_{i2}) = (r_{o0}, r_{o1}, r_{o2}) =: (r_0, r_1, r_2)$$

in this section. The fixed points  $(x_1^*, x_2^*)$  are easily obtained by setting  $(d/dt)(x_1, x_2) = (0, 0)$  in eq. (4) to find  $(x_1^*/x_2^*) = (c_1\Delta_2/c_2\Delta_1)$ , so that

$$c_1\varphi(x_1^*, (c_2\Delta_1/c_1\Delta_2)x_1^*) - \Delta_1x_1^* = 0 \quad (12)$$

is a cubic equation with either 1 or 3 real solutions. Since we can set a rescaled time variable, there are principally 3 parameters that determine the different dynamical outcomes of this model. We shall adjust the ratio of the decay rates  $\Delta = \Delta_1/\Delta_2$  to capture the different time-scales for the two copies, the ratio  $c = c_1/c_2$  which includes the number of binding sites and can be used to model diploid versions of the duplicated haplotype, and  $r = r_1/r_2$  the relative affinities for the activators to the transcriptional machinery.

### 5.1. Oscillations by local and global bifurcations

The stability of a fixed point is usually investigated by the behaviour of the vector-field in its vicinity, *i.e.*, by observing how the system responds to a perturbation about that point. Using the factorisation eq. (10) of the Jacobian of the full system eq. (4) we focus our interest on the two-activator subnetwork as being the generator of novelty rather than the dosage dependence downstream. In particular, we seek out the conditions for a Hopf bifurcation to find oscillatory solutions to the equations (see Appendix C), using Sylvester resultants [50].

There, we find two conditions for a Hopf bifurcation, by factoring a polynomial constraint into two pieces. First,

$$\frac{c}{\Delta} = \sqrt{-\frac{r_2 - r_0}{r_1 - r_0}}$$

which implies that one of the two genes recruits Pol II more efficiently than the basal rate while the other's activation rate is less than that of basal transcription – *i.e.*, one is an activator, the other a repressor. This topology was proposed in [51] and implemented in [52] and shown on the right in Figure 9. We do not deal with this case in this paper.

The second factor in the polynomial constraint obtained by the Sylvester resultant in Appendix C gives rise to the model that we work with in this paper, shown on the left in Figure 9. This factor gives a surface in the 3-dimensional space determined by  $((r_2/r_1), (c_2/c_1), (\Delta_1/\Delta_2))$  as shown in Figure 8 for chosen values of  $\Delta_2$ ,  $c_2$  and  $r_0$ . This demonstrates that a two-activator system can sustain stable oscillations and illustrates how topology is an insufficient predictor of network function. As suggested by the observation of dual regulation by  $x_2$  as a function of  $x_1$  levels, low-expression states participate in a positive feedback loop that maintain a stable steady state, whereas for high  $x_1$  levels  $x_2$  behaves like a repressor and the oscillatory behaviour of positive and negative loops [51] is observed.

The subset of the parameter ranges for which the system oscillates in Figure 11 (a) where the light grey and dark grey (red online) regions show different kinds of stable dynamics. The dark (red online) region displays a region where there is only one fixed point with stable oscillations as shown in Figure 11 (b) even for a stochastic version of the model (see below). The light grey region is where a stable limit cycle and a steady state coexist. To give an indication of the biological plausibility of such a mechanism, the biophysical parameter of interest is the ratio of binding rates to the transcriptional machinery,  $r = r_1/r_2$ . To generate oscillations, a value around  $r = 50$  is sufficient for small values of  $c$ . This translates into a free-energy of binding differential of  $\Delta\Delta G \sim 2.3$  kcal/mol.

We also point out that there are global bifurcations in this model that lead to the emergence of oscillations upon parameter changes, as shown in Figure 10.

### 5.2. Coexistence of one stable equilibrium and one oscillatory state

The presence of a saddle node separating a stable steady state and an unstable fixed point with complex eigenvalues enables the system to be at either of these two states. Below the threshold set by the saddle, the system settles into the stable steady state; above it, the system oscillates. A saddle-node bifurcation arises when the stable fixed point and the saddle merge [45] (the light-dark (grey-red online) boundary in Figure 11(a)), leaving the system to exist only its oscillatory state.

The coexistence of an equilibrium state and a stable limit cycle (the light (grey) volume shown in Figure 11) enables oscillations to be annihilated by the suitably chosen perturbation [53] as well as for rhythms to be switched back on, but with a reset phase, by a threshold-crossing perturbation around the stable fixed point. Such a mechanism underlies a temperature compensa-



tion scheme in circadian clocks in flies [54], suggesting plausibility arguments for fitness parameters that favour such a phenotype.

### 5.3. *Noise-driven oscillations and bursts*

The coexistence of stable equilibrium and stable limit cycle dynamics in the deterministic model suggests that a stochastic version of the model, incorporating intrinsic noise of biochemical reaction steps that constitute the network, will show switching behaviour between these phenotypes. The stochastic kinetic model (which is a simplified reduction of the detailed set of molecular reactions presented in Appendix A.3 ) contains a “volume parameter”  $\Omega$  which enables the translation from (nano-)molar concentrations used in deterministic chemical kinetics models implemented as ODEs and numbers of molecules that are present in cellular volumes. We have generated sample paths using Gillespie’s algorithm [55] for different values of  $\Omega$  to observe the extent of noise in the trajectories as well as the frequency of switching between these states. We illustrate the occurrence of noise driven transitions in Figure 11(c,d). By modulating the threshold by an external input it would also be possible to generate frequency modulated bursts of oscillations, an additional mode of dynamical processing made available to the cell via such a double-activator subnetwork.

### 5.4. *Population levels are at steady state even when cells oscillate*

The bursty oscillatory character of the system reflects noise-driven switching between a stable steady state and a high mean-expression oscillatory state. The random nature of the switching resets the phase and the lack of coherence in an unsynchronised population between individual oscillators results in very low amplitude oscillations. This averaged behaviour is not significantly different from a steady state population response. Furthermore, univariate histograms for the system that displays stable oscillations or oscillatory bursts display bimodal distributions of proteins reminiscent of switches. Thus, it is plausible that a population of cells could present a responsive interface to environmental inputs, such as driving metabolic processing of nutrients in a manner that is neutral, and hides the novel time-dependence in individual cells. This might enable anticipatory responses to periodic environmental cues by processes further downstream to develop and manifest themselves, or indeed be stabilised via duplication events in diploid species (see below). The smaller parameter range shown in Figure 11 when the steady state is lost and only the oscillatory state remains indicates that it is possible for the system to mask the potentially harmful acquisition of

oscillatory instability and commit to oscillations under the future favourable environmental conditions for its subsequent selection.

## 6. Evolution of duplicates for the oscillating phenotype

The outcome of introducing a duplicate mutant will be to either successfully invade the resident population of singleton alleles or die out. In this paper we do not explicitly model the effects of drift, but we note that in small populations, selection coefficients much smaller than  $(1/N_e)$ , where  $N_e$  is the effective population size will get swamped by drift. Moreover, while there were two scenarios explored in the dynamics of regulation – a switch-like or an oscillatory phenotype – we shall only explore the evolutionary consequences of oscillatory dynamics and its role in the spread of the duplicate gene in this paper. Recall, the fitness coefficients in (2) are organised to explore the consequences of increased fitness arising from unequal activation strengths of the two alleles of a self-activating gene upon the potential for retaining gene duplicates. In the singleton case, homozygotes  $a_1/b_1$  or  $a_2/b_2$  (note:  $a_2$  and  $b_2$  are the same allele) have  $r_1 = r_2$  and  $c_1 = c_2$ , and hence the phenotype is non-oscillating. The allocation of fitness contributions  $(1 - s)$  and  $(1 - t)$  imply that there is positive selection for the heterozygote singleton. While this may arise for a variety of factors, we have identified the onset of oscillations in both deterministic and stochastic versions of the regulatory network model as the main qualitative difference. Hence, we investigate the potential for the duplicate to increase in frequency under conditions that do not require additional fitness advantages to doubling  $c := c_1/c_2$  (5), setting  $d = 0$  in (2). (One can see in Figure 11 that increasing  $c = c_1/c_2$  pushes the system deeper into the oscillatory region, and indeed, can drive a Hopf bifurcation, as seen in Figure 8.) One question we investigate here is whether halving  $c$  in (5) in the genotypes  $a_1b_0/a_2b_2$ ,  $a_2b_0/a_1b_2$  and  $a_2b_0/a_2b_1$ , which has the phenotypic effect of reducing the propensity for oscillations, affects the potential for the duplicate gene to spread. For this, we set  $u = t + u'$ , where  $u' = 0$  is the condition that it matches the fitness of homozygote singleton  $a_2b_2$  genotype, and do a perturbative analysis around  $u = t$ .

Since both 2-by-2 recombination matrices in (8) are positive, their largest eigenvalue has a unique corresponding eigenvector (called the Perron eigenvector) which has all entries of the same sign (taken positive), by the Perron-Frobenius theorem. If the corresponding eigenvalue is greater than 1, the linear combination of haplotypes defined by the entries of the Perron eigenvector increases in frequency due to the process of recombination with the existing singleton haplotypes in equilibrium.

The  $\mathbf{R}_{11 \times 21}$  recombination map has a largest eigenpair  $(\lambda, \mathbf{v})$  to first order in  $(t - u)$  and  $d$ :

$$\begin{aligned}\lambda - 1 &= \frac{s^2}{\hat{W}(s+t)^2} \left( (t-u) + d \left( \frac{t}{s} \right) + \dots \right) \\ \mathbf{v} &= \begin{pmatrix} \frac{t}{s} - (t-u) \frac{t}{\rho(s+t)} - d \frac{t(t-s)}{\rho s(t+s)} + \dots \\ 1 \end{pmatrix}. \end{aligned} \quad (13)$$

This shows the relative evolutionary consequences of dosage *balance* between the alternative alleles  $a_1$  and  $a_2$  by noting the unequal contributions of  $[\phi_1] = 2[\phi_2]$  and  $2[\phi_1] = [\phi_2]$  to fitness. If the phenotype of oscillations is favoured, we have  $[\phi_1] = 2[\phi_2]$ , whereby  $W_{10;21} = W_{10;12} = W_{20;11} = 1 + d$  with  $d \geq 0$ ;  $d > 0$  indicates the positive benefits of dosage imbalance and thereby increasing the amplitude of oscillations (data on amplitudes not shown). If  $2[\phi_1] = [\phi_2]$ ,  $c \rightarrow (c/2)$  upon duplication, which reduces the likelihood of oscillations (see Figure 8). If the value of  $c$  were sufficiently high preceding duplication (due to the various factors in eq. (5)), this might yet permit the system to retain an oscillatory phenotype. leading to the condition  $t > u$ . Hence, eq. (13) shows that no additional benefit has to accrue to the oscillatory phenotype ( $d = 0$ ) for the duplicate to propagate if pre-duplication allelic divergence is selected for [21; 17; 5]. The high-amplitude oscillatory state where  $[\phi_1] = 2[\phi_2]$  may indeed be weakly selected against ( $d < 0$ ) and yet enable the spread of the duplicate. Consequently, the proportion of the duplicated allele  $a_1 b_1$  is reduced compared to  $a_2 b_1$  as evident from the Perron eigenvector. This reduction is larger for tightly linked genes (small  $\rho$ ,  $0 \leq \rho \leq (1/2)$ ). This suggests different consequences for duplicated genes whether they arise by tandem duplication or by retrotransposition.

## 7. Discussion

In this paper, we find that duplication of an autoregulatory gene can lead to competition for common genomic binding sites. Assuming a cooperative mode of regulation described via, but not requiring, a dimeric mechanism, we find that a dual regulatory behaviour – context dependent activation or repression – ensues. Feedback amplification of dual regulation gives rise to switching and oscillatory behaviour. While a minimal model like the one we have chosen involves an autoregulatory loop, we anticipate similar qualitative changes to emerge in the more general context of duplication of the node closing a larger feedback loop. The model displays mutually

exclusive expression levels under appropriate contextual cues, analogous to those specified by the subfunctionalization model for retention of duplicates. When the 2-node double-activator subnetwork is viewed as alleles in a diploid species, allelic divergence of expression emerges in this case. We note, in this context, that partitioning of expression of different alleles occurs in  $F_1$  hybrid cotton plants for the alcohol dehydrogenase *Adh* gene [56]. Moreover, we also find that the ensuing circuit is capable of back-up of expression of a deleted gene by its paralog kafri-pnas06. This is linked to developmental systems where such a phenomenon has been observed. The evolutionary instability of this genetic buffering has not been modelled, however, as we have not included the effects of mutational loss in our evolutionary analysis.

We have constructed fitness functions that favours the heterozygote, implying that coupling to periodicity in the environment enhances the selective advantage of an organism bearing such a duplicate. In an *in silico* evolution experiment it has been shown that in a periodic environment increased metabolic flux is obtained via glycolytic oscillations when compared to a non-oscillating response [57]. Further, it has been experimentally realised that hexose transporter genes regulating influx of sugar is crucial for the establishment of yeast glycolytic oscillations [58], and yeast evolves to outcompete ancestral colonies in nutrient-limited conditions by duplicating hexose transporter genes including *HXT6/HXT7* [59]. The coupling of oscillatory phenotype to environmental cues has been most often discussed in the context of circadian rhythms [60; 61]. The results of [22] show that increased biomass yields of hybrids [8] –  $F_1$  (allopolyploid) crosses of two *Arabidopsis* species – was correlated with increases in amplitude of circadian rhythms of clock and clock-controlled genes. While  $F_1$  hybrids are often evolutionarily unstable, our simple evolutionary model shows the spread of a duplicate in such a system with overdominance. Circadian rhythm generating mechanisms have been much studied, and it has been shown in mammalian systems that competitive binding to common conserved *cis*-regulatory regions (such as the E/E'-box, the RRE and the D-box in mammals [62]) drives circadian rhythms, and that its removal reduces the amplitude of oscillations. While this appears to be consonant with our findings, it should be pointed out competitive binding is the sole mechanism in our model, unlike in the circadian clock. The competitive mechanism for the generation of oscillations relies on the increased efficacy of activation of one of the activators feeds greater proteolytic rates [48; 63]. The opposite conclusion to "competition aids oscillations" is drawn in numerical analysis of models of synthetic oscillators [64].

Our model displays the role of copy numbers of alleles in increasing the

amplitude of oscillations (via increasing  $c$ ). Consequences of gene duplication have been considered in the context of dosage balance. The ubiquity of dominance has given rise to the hypothesis that dosage balance is favoured [8]. On the other hand, in the case of the oscillatory dynamical states discussed, our assumptions favoured a heterozygote – that under some environmental (typically periodic) conditions there is selective advantage to oscillatory behaviour – leading to the increase in frequency of gene duplicates with genotypes with an even larger amplitude for oscillations. On a large-scale analysis, it has been noted that there is an increased fraction [25] of duplicated yeast genes as a result of whole genome duplication amongst those that cycle during metabolic oscillations [65], and [26] identify the extensive participation of paralogs in multiple rhythmic processes as a partitioning of the oscillatory feature amongst duplicates. We leave for a future investigation the theoretical analysis of duplications in such regulatory model systems.

## Appendix A. Model construction

In this Appendix we describe, first, the thermodynamic model of transcription rates, then the detailed kinetic model that facilitates the introduction of the stochastic model used for performing the simulations. We also introduce the context dependent model which takes the influence of *cis*-regulatory site information in the rate of transcription.

### Appendix A.1. Thermodynamic model of promoter occupancy

Transcription factors bound to the enhancer recruit RNA polymerase and the thermodynamic formalism can easily accommodate different binding free energies of protein-DNA and protein-protein interactions in a uniform manner. In this paper, we summarize the composite effects of transcription factor-Mediator and Mediator-RNA polymerase binding by energy terms  $\epsilon_{A_i P}$  for each activator  $A_i$ . Hereafter, we denote by  $P$  the transcriptional apparatus involving general transcription factors, the Mediator complex and RNA Pol II. In the standard way to count different possible configurations (see [33] for a pedagogical introduction in the context of gene regulation) we introduce Boltzmann factors for all possible configurations for binding of activators  $A_{1,2}$  and  $P$  to calculate the partition function  $Z_{tot}(P, A_1, A_2)$ . A subset of these configurations are poised for transcription *i.e.*, those with RNA polymerase or Pol II bound to the promoter. The rate of mRNA synthesis is taken to be proportional to the probability of the Pol II bound promoter, which is taken to be the ratio of the Boltzmann

factors for the favourable configurations with promoter-specific bound Pol II to  $Z_{tot}(P, A_1, A_2)$ .

The binding energies for non-specific (site-specific) binding are denoted  $\varepsilon^0(\varepsilon^s)$ , with appropriate subscripts which identify the binding of activators or the polymerases to the DNA. For the case where  $k$  polymerases,  $l$  activators of type 1 and  $m$  activators of type 2 bind to the *cis*-regulatory region of the DNA their energy contribution is

$$E^s(k, l, m) := k(\varepsilon_{pd}^s + l\varepsilon_{A_1p} + m\varepsilon_{A_2p} + lm\varepsilon_{A_1A_2p}) + l\varepsilon_{A_1d}^s + m\varepsilon_{A_2d}^s + lm\varepsilon_{A_1A_2}, \quad (\text{A.1})$$

where the subscripts indicate the protein-protein binding energies as well, including  $\varepsilon_{A_1A_2p}$  which captures the net energy of recruitment of transcriptional machinery due to the combined action of the activators. If the proteins bind to non-cognate sites, the binding energy is

$$E^0(k, l, m) := k\varepsilon_{pd}^0 + l\varepsilon_{A_1d}^0 + m\varepsilon_{A_2d}^0. \quad (\text{A.2})$$

For  $P$  the number of Pol II molecules,  $A_{1,2}$  the number of transcription activators of each type, we introduce

$$\zeta(P, A_1, A_2) = \binom{N_{ns}}{P, A_1, A_2} e^{-\beta E^0(P, A_1, A_2)}, \quad (\text{A.3})$$

where the right hand side includes in the exponent  $\beta = 1/(k_B T)$ , and the trinomial coefficient contains the number of binding sites in the genome  $N_{ns}$ . We shall simplify the partition function for 0 or 1 molecules of type  $P, A_{1,2}$  bound to the relevant DNA sites

$$Z_{tot}(P, A_1, A_2) = \sum_{(k, l, m) \in \{0, 1\}^3} \zeta(P - k, A_1 - l, A_2 - m) e^{-\beta E^s(k, l, m)} \quad (\text{A.4})$$

using the (Stirling) approximation

$$\binom{N_{ns}}{X, Y, Z} \approx \frac{N_{ns}^{X+Y+Z}}{X!Y!Z!} \text{ if } N_{ns} \gg X, Y, Z, \quad (\text{A.5})$$

and the following definitions (A.6),

$$\begin{aligned} \alpha_i &= \frac{A_i}{N_{ns}} e^{-\beta(\varepsilon_{A_i d}^s - \varepsilon_{A_i d}^0)}, \quad r_i = e^{-\beta \varepsilon_{A_i p}}, \quad i = 1, 2 \\ \rho &= \frac{P}{N_{ns}} e^{-\beta(\varepsilon_{pd}^s - \varepsilon_{pd}^0)}, \quad r_{12} = e^{-\beta \varepsilon_{A_1 A_2 p}}, \quad \omega_{12} = e^{-\beta \varepsilon_{A_1 A_2}}. \end{aligned} \quad (\text{A.6})$$

We obtain

$$\frac{Z_{tot}(P, A_1, A_2)}{\zeta(P, A_1, A_2)} = 1 + \rho + (1 + \rho r_1)\alpha_1 + (1 + \rho r_2)\alpha_2 + (1 + \rho r_{12})\omega_{12}\alpha_1\alpha_2, \quad (\text{A.7})$$

which enables us to compute the probability of occupancy of the promoter by RNA Pol II:

$$\frac{1}{Z_{tot}(P, A_1, A_2)} \sum_{(l,m) \in \{0,1\}^2} \zeta(P-1, A_1-l, A_2-m) e^{-\beta E^s(1,l,m)}. \quad (\text{A.8})$$

Finally, we can express the transcription rate as proportional to probability of promoter occupancy by  $P$ :

$$\Phi = \frac{\rho(1 + r_1\alpha_1 + r_2\alpha_2 + r_{12}\omega_{12}\alpha_1\alpha_2)}{(1 + \alpha_1 + \alpha_2 + \omega_{12}\alpha_1\alpha_2) + \rho(1 + r_1\alpha_1 + r_2\alpha_2 + r_{12}\omega_{12}\alpha_1\alpha_2)}. \quad (\text{A.9})$$

Phenomenologically, such an expression is matched to experimental data on amplification or reduction of mRNA production as a function of transcription factor numbers, called the fold change function,  $\psi$ :

$$\psi(\alpha_1, \alpha_2) = \frac{\rho(1 + r_1\alpha_1 + r_2\alpha_2 + r_{12}\omega_{12}\alpha_1\alpha_2)}{(1 + \alpha_1 + \alpha_2 + \omega_{12}\alpha_1\alpha_2)}. \quad (\text{A.10})$$

For the case of competitive binding of  $A_1$  and  $A_2$  to the same promoter site, we shall set  $\omega_{12} = 0$  hereafter.

#### Appendix A.2. Context-dependent recruitment

In order to introduce *cis*-context dependent regulation beyond that captured by  $A_{1,2}$  binding sites, we will need to introduce additional binding sites. We introduce two additional factors  $A_3$  and  $A_4$  but we assume that it only alters the transcriptional ability of  $A_{1,2}$  by site-specific and protein-protein interactions. We proceed in exactly the same lines as before to end up with

$$\psi(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \frac{\rho \left( 1 + \sum_{i=1}^4 r_i \alpha_i + \sum_{k=3}^4 r_{ik} \omega_{ik} \alpha_i \alpha_k \right)}{\left( 1 + \sum_{i=1}^4 \alpha_i + \sum_{k=3}^4 \omega_{ik} \alpha_i \alpha_k \right)}, \quad (\text{A.11})$$

and the promoter occupancy is, as before,

$$\Phi = \frac{1}{1 + \psi^{-1}}.$$

In what follows we shall restrict ourselves to the situation where  $A_3$  and  $A_4$  does not initiate transcription on its own but merely acts as helpers to  $A_{1,2}$ . Thus we can set  $r_3 = 0 = r_4$  and end up with a context dependent fold-change factor  $\psi^{(c)}$

$$\psi^{(c)}(\alpha_1, \alpha_2) = \frac{\rho(1 + \alpha_1(r_1 + r_{13}\omega_{13}\alpha_3 + r_{14}\omega_{14}\alpha_4) + \alpha_2(r_2 + r_{23}\omega_{23}\alpha_3 + r_{24}\omega_{24}\alpha_4))}{(1 + \alpha_1(1 + \omega_{13}\alpha_3 + \omega_{14}\alpha_4) + \alpha_2(1 + \omega_{23}\alpha_3 + \omega_{24}\alpha_4))}, \quad (\text{A.12})$$

which enables a greater flexibility in setting differential rates of recruitment based on both protein-protein interactions  $\omega_{ij}$ , sequence-dependent transcription factor-DNA binding ( $\alpha_i$ ) as well as recruitment of the transcriptional machinery ( $r_i, r_{ij}$ ).

### Appendix A.3. Kinetic model and reduction

In order to perform stochastic simulations it is convenient to also introduce a detailed kinetic scheme which reduces to the result derived from thermodynamics under detailed balance of the fast reactions to be indicated below. As in the thermodynamic description, we shall describe the details of the two activator case and extend the formalism to include the third auxiliary transcription factor as that does not play a dynamical role in the paper, but merely motivates the parameterization.

For this section it is convenient to introduce  $a$  and  $b$  as labels that refer to the two genes, and  $A$  and  $B$  to denote the two proteins that are expressed. We shall also refer to the quantities of  $A$  and  $B$ -monomers by  $[A]$  and  $[B]$  and their dimeric forms by  $[A2]$  and  $[B2]$  respectively. Later on we shall also introduce the variables  $x$  and  $y$  which are scaled versions of  $[A]$  and  $[B]$ . To match the  $(x_1, x_2)$  variables in the paper, we note that  $x = x_1$  and  $y = x_2$ . To connect to the thermodynamic description in A.9,  $\alpha_{1,2}$  will be taken to be dimers  $[A2]$ ,  $[B2]$  and will be represented by  $x^2$  and  $y^2$  in the Appendix , and to repeat, is called  $x_{1,2}$  in the main paper.

The set of reactions that we consider are given in the table below. Note that the reactions below the horizontal line in the table are *excluded*, and are presented only to indicate which reactions occur with vanishing probability because of the steric inhibition of the transcription factors. In order to relate the kinetic description under detailed balance to the thermodynamic description, we set

$$\omega_{12} = 0 \quad (\text{A.13})$$

in A.9 to impose the mutual exclusivity of binding to the enhancers.



Reactions	Rates	Propensities
$2A \rightleftharpoons A_2$	$(k_{dim,A}^b, k_{dim,A}^u)$	$(k_{dim,A}^b n_A^2 / \Omega, k_{dim,A}^u n_{A_2})$
$2B \rightleftharpoons B_2$	$(k_{dim,B}^b, k_{dim,B}^u)$	$(k_{dim,B}^b n_B^2 / \Omega, k_{dim,B}^u n_{B_2})$
$A_2 + DA \rightleftharpoons DAA_2$	$(k_{da,A}^b, k_{da,A}^u)$	$(k_{da,A}^b n_{A_2} n_{DA} / \Omega, k_{da,A}^u n_{DAA_2})$
$A_2 + DB \rightleftharpoons DBA_2$	$(k_{db,A}^b, k_{db,A}^u)$	$(k_{db,A}^b n_{A_2} n_{DB} / \Omega, k_{db,A}^u n_{DBA_2})$
$B_2 + DA \rightleftharpoons DAB_2$	$(k_{da,B}^b, k_{da,B}^u)$	$(k_{da,B}^b n_{B_2} n_{DA} / \Omega, k_{da,B}^u n_{DAB_2})$
$B_2 + DB \rightleftharpoons DBB_2$	$(k_{db,B}^b, k_{db,B}^u)$	$(k_{db,B}^b n_{B_2} n_{DB} / \Omega, k_{db,B}^u n_{DBB_2})$
$DA + Pol \rightleftharpoons DAPol$	$(\rho_{a0}^+, \rho_{a0}^-)$	$(\rho_{a0}^+ n_{DA} n_{Pol} / \Omega, \rho_{a0}^- n_{DAPol})$
$DB + Pol \rightleftharpoons DBPol$	$(\rho_{b0}^+, \rho_{b0}^-)$	$(\rho_{b0}^+ n_{DB} n_{Pol} / \Omega, \rho_{b0}^- n_{DBPol})$
$DAA_2 + Pol \rightleftharpoons CAA_2$	$(\rho_{aA}^+, \rho_{aA}^-)$	$(\rho_{aA}^+ n_{DAA_2} n_{Pol} / \Omega, \rho_{aA}^- n_{CAA_2})$
$DAB_2 + Pol \rightleftharpoons CAB_2$	$(\rho_{aB}^+, \rho_{aB}^-)$	$(\rho_{aB}^+ n_{DAB_2} n_{Pol} / \Omega, \rho_{aB}^- n_{CAB_2})$
$DBA_2 + Pol \rightleftharpoons CBA_2$	$(\rho_{bA}^+, \rho_{bA}^-)$	$(\rho_{bA}^+ n_{DBA_2} n_{Pol} / \Omega, \rho_{bA}^- n_{CBA_2})$
$DBB_2 + Pol \rightleftharpoons CBB_2$	$(\rho_{bB}^+, \rho_{bB}^-)$	$(\rho_{bB}^+ n_{DBB_2} n_{Pol} / \Omega, \rho_{bB}^- n_{CBB_2})$
$DAPol \rightarrow DA + m_a + Pol$	$\mu_{a,0}$	$\mu_{a,0}$
$DBPol \rightarrow DB + m_b + Pol$	$\mu_{b,0}$	$\mu_{b,0}$
$CAA_2 \rightarrow DAA_2 + m_a + Pol$	$\mu_{a,A}$	$\mu_{a,A}$
$CAB_2 \rightarrow DAB_2 + m_a + Pol$	$\mu_{a,B}$	$\mu_{a,B}$
$CBA_2 \rightarrow DBA_2 + m_b + Pol$	$\mu_{b,A}$	$\mu_{b,A}$
$CBB_2 \rightarrow DBB_2 + m_b + Pol$	$\mu_{b,B}$	$\mu_{b,B}$
$DAPol \rightarrow DA + m_a + Pol$	$\mu_{a,0}$	$\mu_{a,0}$
$DBPol \rightarrow DB + m_b + Pol$	$\mu_{b,0}$	$\mu_{b,0}$
$m_a \rightarrow A + m_a$	$\pi_a$	$\pi_a$
$m_a \rightarrow \emptyset$	$\delta_{ma}$	$\delta_{ma}$
$m_b \rightarrow B + m_b$	$\pi_b$	$\pi_b$
$m_b \rightarrow \emptyset$	$\delta_{mb}$	$\delta_{mb}$
$A \rightarrow \emptyset$	$\Delta_A$	$\Delta_A$
$B \rightarrow \emptyset$	$\Delta_B$	$\Delta_B$
$B_2 + DAA_2 \rightleftharpoons DAX$	$(k_{da,X}^b, k_{da,X}^u) = (0, \infty)$	$(0, \infty)$
$A_2 + DBB_2 \rightleftharpoons DBX$	$(k_{db,X}^b, k_{db,X}^u) = (0, \infty)$	$(0, \infty)$

Table A.3: The molecular species/states represented in the reaction scheme are labelled by the genetic identities,  $A$  and  $B$ .  $DA, DB$  stand for promoter regions upstream of genes  $a, b$ . mRNA and proteins are  $m_a, A$ , etc, while Pol is a shorthand for the set of intermediates including Mediator and the RNA Pol II transcriptional machinery. Of key significance in this paper is the different affinities of the transcription factors  $A_2$ , etc bound to promoters  $DAA_2$ , etc have for this transcribing machinery. The states  $DAA_2$  and  $DAB_2$  refer to the states of the promoter of gene  $a$  bound by  $A_2, B_2$  respectively. The set of reactions below the horizontal line involving states DAX, CAX, DBX, CBX are those with promoters bound by both transcription factors. These states are excluded in the XOR case,  $k_{da,X}^b = 0 = k_{db,X}^b$ . We have also set the basal rates of polymerase binding to the two promoters the same,  $\rho_0^\pm$ . The rightmost column gives the propensities for the reactions used in the Gillespie simulation. The probabilities for the stochastic case are obtained by dividing the rate constants by  $N_A \Omega$  where  $N_A$  is Avogadro's number and  $\Omega$  a volume. Under this normalisation, a 1nM concentration corresponds to approximately 1 molecule in *E. coli* and 60 molecules in a mammalian cell nucleus.

In the reactions in Table A.3, we make the assumption that all the binding-unbinding events in the *cis*-regulatory regions are much faster compared to the slow processes of transcript formation and translation. Further, we assume detailed balance to arrive at the same fractions for the states corresponding to the bound configurations CAA2, CAB2, CBA2, CBB2 (and CA, CB for basal transcription) as in the thermodynamical formulation A.9.

The rates of transcription of the mRNA species of A and B are:

$$\begin{aligned}\frac{d}{dt}[m_a] &= \mu_{a,0}[CA] + \mu_{a,A}[CAA2] + \mu_{a,B}[CAB2] - \delta_a[m_a] \\ \frac{d}{dt}[m_b] &= \mu_{b,0}[CB] + \mu_{b,A}[CBA2] + \mu_{b,B}[CBB2] - \delta_b[m_b]\end{aligned}\quad (\text{A.14})$$

where the promoters [DA], [DB] of genes A and B are bound by the RNA Pol II and initiate transcript elongation at the different rates  $\mu$ . We shall make the assumption in this paper that transcription elongation takes place at a rate that is independent of promoter configuration and so we shall set

$$\mu_{a,0} = \mu_{a,A} = \mu_{a,B} = \mu_a \text{ and } \mu_{b,0} = \mu_{b,A} = \mu_{b,B} = \mu_b.$$

In order to match up with the thermodynamic formalism we need to *assume* detailed balance for all the reactions where  $\triangleright$  binds to/unbinds from  $\triangleleft$  to form  $\bowtie$  in Table A.3:

$$k^+[\triangleright][\triangleleft] = k^-[\bowtie] \Rightarrow [\bowtie] = \frac{k^+}{k^-}[\triangleright][\triangleleft], \quad (\text{A.15})$$

which prompts us to introduce:

$$K_{dim,A} = \frac{k_{dim,A}^u}{k_{dim,A}^b}, \quad K_{dim,B} = \frac{k_{dim,B}^u}{k_{dim,B}^b} \quad (\text{A.16})$$

for the dimer dissociation constants and

$$K_{da,A} = \frac{k_{da,A}^u}{k_{da,A}^b}, \quad K_{db,A} = \frac{k_{db,A}^u}{k_{db,A}^b}, \quad K_{da,B} = \frac{k_{da,B}^u}{k_{da,B}^b}, \quad K_{db,B} = \frac{k_{db,B}^u}{k_{db,B}^b} \quad (\text{A.17})$$

for protein-DNA binding, where the  $K_{d\circ,\bullet}$  notation stands for  $\bullet$  binding to promoter of gene  $\circ$ .

The GRF involves terms of the form  $([TF]/K_{TF}) \times r$  where  $r$  incorporates the binding probabilities of the transcription factors to the transcriptional machinery (such as Mediator, RNA polymerases, etc) and  $K_{TF}$  is the dissociation constant for protein-DNA interaction (involves exponentials of free

energy of binding). The TFs of interest in this model are dimers  $A_2$  of protein  $A$  and from

$$A + A \xrightleftharpoons[k_{dim,A}^u]{k_{dim,A}^b} A_2 \quad K_{dim}^A = \frac{k_{dim,A}^u}{k_{dim,A}^b} = \frac{[A]^2}{[A_2]}$$

. Using the detailed balance rapid equilibrium hypothesis, the dynamical variables enter with products of the dissociation constants for protein-DNA and dimerisation dissociation constants.

We introduce

$$r_{i0} := [\text{Pol}] \frac{\rho_{i0}^+}{\rho_{i0}}, \quad r_{iA} := [\text{Pol}] \frac{\rho_{iA}^+}{\rho_{iA}}, \quad r_{iB} := [\text{Pol}] \frac{\rho_{iB}^+}{\rho_{iB}} \quad (\text{A.18})$$

where  $i$  could take the index values  $a$  or  $b$  and would reflect the promoter dependence of the (basal) dissociation rates in question. We further define  $\kappa_A^2 = K_{dim}^A K_{da,A}$  and  $\kappa_B^2 = K_{dim}^B K_{db,B}$  so that

$$x = [A]/\kappa_A \quad \text{and} \quad y = [B]/\kappa_B,$$

and the ratios  $t_{ij}^{-1} = (K_{di,j}/K_{dj,j})$ , with  $t_{ii} = 1$ . Thus,  $t_{aB}^{-1} = (K_{da,B}/K_{db,B})$ ,  $t_{bA}^{-1} = (K_{db,A}/K_{da,A})$  measure the relative strengths of binding of the transcription factors to the promoter regions of the 2 genes using their autoregulatory binding as reference. Now we substitute for the bound promoter concentrations in eq. (A.14) using the relations derived from detailed balance to obtain

$$\begin{aligned} \frac{d}{dt}[m_a] &= -\delta_a[m_a] + \mu_a[DA][\text{Pol}] \left( \frac{\rho_{a0}^+}{\rho_{a0}} + \frac{\rho_{aA}^+}{\rho_{aA}} \frac{[A]^2}{K_{dim}^A K_{da,A}} + \frac{\rho_{aB}^+}{\rho_{aB}} \frac{K_{db,B}}{K_{da,B}} \frac{[B]^2}{K_{dim}^B K_{db,B}} \right) \\ &= -\delta_a[m_a] + \mu_a[DA] (r_{a0} + r_{aA}x^2 + r_{aB}t_{aB}y^2) \\ &\quad (\text{A.19}) \\ \frac{d}{dt}[m_b] &= -\delta_b[m_b] + \mu_b[DB][\text{Pol}] \left( \frac{\rho_{b0}^+}{\rho_{b0}} + \frac{\rho_{bA}^+}{\rho_{bA}} \frac{K_{da,A}}{K_{db,A}} \frac{[A]^2}{K_{dim}^A K_{da,A}} + \frac{\rho_{bB}^+}{\rho_{bB}} \frac{[B]^2}{K_{dim}^B K_{db,B}} \right) \\ &= -\delta_b[m_b] + \mu_b[DB] (r_{b0} + r_{bA}t_{bA}x^2 + r_{bB}y^2). \end{aligned}$$

To obtain the expressions for the promoter fractions  $[DA]$  and  $[DB]$ , note that the total available promoters are either unoccupied, or occupied by

transcription factors and core transcriptional machinery. Thus, the total promoter availability is used to obtain the bound fractions, in line with the thermodynamic description above. These are obtained as follows:

$$\begin{aligned}
[DA]_0 &= [DA] + [CA] + [DAA2] + [CAA2] + [DAB2] + [CAB2] \\
\Rightarrow \frac{[DA]_0}{[DA]} &= \left(1 + \frac{\rho_{a0}^+}{\rho_{a0}^-} [\text{Pol}]\right) + \left(1 + \frac{\rho_{aA}^+}{\rho_{aA}^-} [\text{Pol}]\right) \frac{[A]^2}{K_{dim}^A K_{da,A}} \\
&\quad + \left(1 + \frac{\rho_{aB}^+}{\rho_{aB}^-} [\text{Pol}]\right) \frac{K_{db,B}}{K_{da,B}} \frac{[B]^2}{K_{dim}^B K_{db,B}} \\
&= (1 + r_{a0}) + (1 + r_{aA})x^2 + (1 + r_{aB})t_{aB}y^2 \equiv \Gamma_a \\
[DB]_0 &= [DB] + [CB] + [DBA2] + [CBA2] + [DBB2] + [CBB2] \\
\Rightarrow \frac{[DB]_0}{[DB]} &= \left(1 + \frac{\rho_{b0}^+}{\rho_{b0}^-} [\text{Pol}]\right) + \left(1 + \frac{\rho_{bA}^+}{\rho_{bA}^-} [\text{Pol}]\right) \frac{K_{da,A}}{K_{db,A}} \frac{[A]^2}{K_{dim}^A K_{da,A}} \\
&\quad + \left(1 + \frac{\rho_{bB}^+}{\rho_{bB}^-} [\text{Pol}]\right) \frac{[B]^2}{K_{dim}^B K_{db,B}} \\
&= (1 + r_{b0}) + (1 + r_{bA})t_{bA}x^2 + (1 + r_{bB})y^2 \equiv \Gamma_b
\end{aligned} \tag{A.20}$$

We now write the set of equations determining the kinetics of transcription and translation:

$$\begin{aligned}
\frac{d}{dt}[m_a] &= \mu_a[DA]_0 \frac{r_{a0} + r_{aA}x^2 + r_{aB}t_{aB}^{-1}y^2}{(1 + r_{a0}) + (1 + r_{aA})x^2 + (1 + r_{aB})t_{aB}y^2} - \delta_a[m_a], \\
\frac{d}{dt}[A] &= \pi_a[m_a] - \Delta_A[A], \\
\frac{d}{dt}[m_b] &= \mu_b[DB]_0 \frac{r_{b0} + r_{bA}t_{bA}x^2 + r_{bB}y^2}{(1 + r_{b0}) + (1 + r_{bA})t_{bA}x^2 + (1 + r_{bB})y^2} - \delta_b[m_b], \\
\frac{d}{dt}[B] &= \pi_b[m_b] - \Delta_B[B].
\end{aligned} \tag{A.21}$$

We then introduce the assumption that the fast decay times of mRNA compared to those of proteins enables the translation machinery to effectively see a steady state level of mRNA  $[m_i]^{ss}$

$$[m_i]^{ss} = \frac{1}{\delta_i} \text{mRNA production rate}(i), \text{ for } i = a, b \tag{A.22}$$

which we substitute in the differential equations for proteins to arrive at

$$\begin{aligned}
\frac{d}{dt}x &= [DA]_0 \frac{\pi_a \mu_a}{\delta_a \kappa_A} \frac{r_{a0} + r_{aA}x^2 + r_{aB}t_{aB}y^2}{(1 + r_{a0}) + (1 + r_{aA})x^2 + (1 + r_{aB})t_{aB}y^2} - \Delta_A x \\
&\equiv c_a \varphi_a(x, y) - \Delta_A x, \\
\frac{d}{dt}y &= [DB]_0 \frac{\pi_b \mu_b}{\delta_b \kappa_B} \frac{r_{b0} + r_{bA}t_{bA}x^2 + r_{bB}y^2}{(1 + r_{b0}) + (1 + r_{bA})t_{bA}x^2 + (1 + r_{bB})y^2} - \Delta_B y \\
&\equiv c_b \varphi_b(x, y) - \Delta_B y,
\end{aligned} \tag{A.23}$$

where  $c_a = \frac{\pi_a \mu_a}{\delta_a \kappa_A}$  and  $c_b = \frac{\pi_b \mu_b}{\delta_b \kappa_B}$ .

To summarise the correspondence to the notation in the main text, we have  $(x_1, x_2) \leftrightarrow (x, y)$ ,  $(t_{aA}, t_{aB}, t_{bA}, t_{bB}) \leftrightarrow (t_{11}, t_{12}, t_{21}, t_{22})$  and  $(r_{aA}, r_{aB}, r_{bA}, r_{bB}) \leftrightarrow (r_{11}, r_{12}, r_{21}, r_{22})$ , and similarly, the  $a$  and  $b$  subscripts correspond to  $_1$  and  $_2$  in the main text.

#### Appendix A.4. Introducing cis-regulatory context and subfunctionalizing mutations

This section deals with the case of two transcriptional activators  $A_1$  and  $A_2$  and two helper proteins,  $C_1$  and  $A_2$ . These helper proteins facilitate discussions on *cis*-regulatory context and parameterize such context-dependence. For simplicity, we shall restrict ourselves to the situation where  $C_k$  does not activate either of the duplicate genes by itself, but modifies the recruitment potential of  $A_1$  and  $A_2$  via protein-protein interactions and DNA binding affinity at the regulatory site, incorporating the roles of *cis*- and *trans*-effects in evolution. Using the fold change function (A.12) as reference, we define

$$\psi_j^{(C)}(x_1, x_2) = \frac{r_{j0} + (r_{j1} + \sum_{k=1}^2 r_{j1C_k} t_{jC_k} x_{C_k}) t_{j1} x_1^2 + (r_{j2} + \sum_{k=1}^2 r_{j2C_k} t_{jC_k} x_{C_k}) t_{j2} x_2^2}{1 + (1 + \sum_{k=1}^2 t_{jC_k} x_{C_k}) t_{j1} x_1^2 + (1 + \sum_{k=1}^2 t_{jC_k} x_{C_k}) t_{j2} x_2^2} \tag{A.24}$$

which incorporates the protein-protein interactions between transcription factors  $A_{1,2}$  and helper proteins  $C_{1,2}$  that are significant for recruitment of the transcription machinery:  $r_{jiC_k}$  stands for the affinity of the  $A_i$ - $C_k$  protein complex on the DNA regulatory region of gene  $j$  to the transcription machinery (Mediator, Pol II, *etc*).  $t_{jC_k}^{-1} = K_{dj, C_k}$  measures the protein-DNA

dissociation constant of  $C_k$  to the enhancer of  $j$ . We introduce

$$\begin{aligned}\tilde{t}_{jk} &\equiv t_{jk}\left(1 + \frac{C_1}{K_{dj,C_1}} + \frac{C_2}{K_{dj,C_2}}\right) = t_{jk}(1 + t_{jC_1}x_{C_1} + t_{jC_2}x_{C_2}) \text{ and} \\ \tilde{r}_{jk} &\equiv \frac{r_{jk} + r_{jkC_1}t_{jC_1}x_{C_1} + r_{jkC_2}t_{jC_2}x_{C_2}}{1 + t_{jC_1}x_{C_1} + t_{jC_2}x_{C_2}},\end{aligned}\tag{A.25}$$

which simplifies the expression for the fold-change to

$$\psi_j^{(C)}(x_1, x_2) = \frac{r_{j0} + \tilde{r}_{j1}\tilde{t}_{j1}x_1^2 + \tilde{r}_{j2}\tilde{t}_{j2}x_2^2}{1 + \tilde{t}_{j1}x_1^2 + \tilde{t}_{j2}x_2^2}.\tag{A.26}$$

To evaluate the nature of the regulatory activity of  $x_1$  and  $x_2$  in the presence of  $C$  we compute the partial derivatives  $\partial_{x_k}\Phi_i$  of the probability of occupancy of the promoter of  $i$ :

$$\Phi_i = \frac{1}{1 + \psi_i^{-1}}.$$

This can be split into two factors:

$$\frac{\partial\Phi_i}{\partial x_k} = \frac{2x_k}{((1 + \psi_i)(1 + \tilde{t}_{i1}x_1^2 + \tilde{t}_{i2}x_2^2))^2} \left( (\tilde{r}_{ik} - r_{i0}) + \sum_{j=1}^2 (1 - \delta_{kj})\tilde{t}_{ij}x_j^2(\tilde{r}_{ik} - \tilde{r}_{ij}) \right),$$

for  $(i, k = 1, 2)$ , where the second factor determines the sign of the regulatory activity. The dependence of  $\tilde{r}$  on protein-protein interactions  $r_{ijC_k}$  and DNA binding strengths  $t_{iC_k}$  illustrates how context-dependent changes in the nature of regulation can be achieved.

In particular, we demonstrate how complementary loss of function mutations can yield the parameters for the exclusive switching circuit. If all protein-protein interactions are modular, in that they occur due to contact interactions, we can set  $r_{ij} = r_{\circ j}$  and  $r_{ijC_k} = r_{\circ jC_k}$ . (We denote the locus independence, or the lack of an index, as  $\circ$ .) We assume that  $r_{\circ 1C_1} = r_{\circ 2C_2} \equiv r_C$  and  $r_{\circ 2C_1} = r_{\circ 1C_2} = \epsilon r_C$ , for  $\epsilon \ll 1$ . Mutations are assumed to leave the protein-protein interactions unaffected for the *cis*-context cases. To implement the subfunctionalization model we set  $t_{1C_1} = t_{2C_2} \equiv t_C$ , and  $t_{1C_2} = t_{2C_1} = \epsilon t_C$ ; thus, two loss of binding site mutations for the helper proteins  $C_{1,2}$  are assumed to take place. These interaction strengths are thus of the same order of magnitude as the binding to non-specific sites and can thus be absorbed into  $r_{i0}$ . Further, we assume both helpers to be of the same concentration, to simplify description and analysis:  $x_{C_1} \approx x_{C_2} = x_C$ .

Reactions	Rates	Propensities
$A2 + DZ \rightleftharpoons DZA2$	$(k_{da,A}^b, k_{dz,A}^u)$	$(k_{dz,A}^b n_{A2} n_{DZ} / \Omega, k_{dz,A}^u n_{DZA2})$
$B2 + DZ \rightleftharpoons DZB2$	$(k_{dz,B}^b, k_{dz,B}^u)$	$(k_{dz,B}^b n_{DZ} n_{B2} / \Omega, k_{dz,B}^u n_{DZB2})$
$DZ + Pol \rightleftharpoons DZPol$	$(\rho_{z0}^+, \rho_{z0}^-)$	$(\rho_{z0}^+ n_{DZ} n_{Pol} / \Omega, \rho_{z0}^- n_{DZPol})$
$DZA2 + Pol \rightleftharpoons CZA2$	$(\rho_{zA}^+, \rho_{zA}^-)$	$(\rho_{zA}^+ n_{DZA2} n_{Pol} / \Omega, \rho_{zA}^- n_{CZA2})$
$DZB2 + Pol \rightleftharpoons CZB2$	$(\rho_{zB}^+, \rho_{zB}^-)$	$(\rho_{zB}^+ n_{DZB2} n_{Pol} / \Omega, \rho_{zB}^- n_{CZB2})$
$DZPol \rightarrow DZ + m_z + Pol$	$\mu_{z,0}$	$\mu_{z,0}$
$CZA2 \rightarrow DZA2 + m_z + Pol$	$\mu_{z,A}$	$\mu_{z,A}$
$CZB2 \rightarrow DZB2 + m_z + Pol$	$\mu_{z,B}$	$\mu_{z,B}$
$DZPol \rightarrow DZ + m_z + Pol$	$\mu_{z,0}$	$\mu_{z,0}$
$m_z \rightarrow Z + m_z$	$\pi_z$	$\pi_z$
$m_z \rightarrow \emptyset$	$\delta_{mz}$	$\delta_{mz}$
$Z \rightarrow \emptyset$	$\Delta_Z$	$\Delta_Z$

Table A.4: The reactions to be added to the Table 1 to incorporate the action of the transcription factors  $A2$ ,  $B2$  on the target gene  $z$ . The notation is analogous to Table 1 as well.

Making these substitutions into eq. (A.25) we get

$$\begin{aligned}
\tilde{t}_{ij} &\approx t_{ij}(1 + t_C x_C) \\
\tilde{r}_{jk} &\approx \frac{r_{okC_j} + r_{okC_j} t_C x_C}{1 + t_C x_C}.
\end{aligned} \tag{A.27}$$

The strengths  $r_{okC_j}$  determine, for the heterozygous switches considered in the paper, how the complementary strengths of protein-protein interactions for recruitment of polymerases are achieved by setting  $r_{o1C_1} > r_{okC_2}$  and  $r_{o2C_2} > r_{o2C_1}$ .

#### Appendix A.5. Duplicated auto-activator on target gene

Here we consider the case where an activator gene  $a$  activates itself, up-regulating the production of protein  $A$  and turns on a gene  $z$  which expresses a protein  $Z$ . As explained in the text, this is the typical motif that figures in selector genes or terminal selector genes [11]. We examine the consequences of duplicating  $a$  so that now we have two copies  $a_1$ ,  $a_2$  which are mutually and self-activating and also inherit a common target site in  $z$ .

If we impose a similar detailed balance condition to extract the kinetic equations from the above reaction scheme, we end up with the following

scheme:

$$\frac{dx_i}{dt} = c_i \frac{1}{1 + \psi_i^{-1}(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i)} - \Delta_i x_i, \quad (\text{A.28})$$

where  $\mathbf{x} = (x_1, x_2)$ , and  $(x_1, x_2, x_3) = ([A_1]/\kappa_1, [A_2]/\kappa_2, [Z])$ , where  $\kappa_i$  is the geometric mean of two dissociation constants  $K_D$ : the dissociation constant of dimerization of protein labelled by  $i$  and the protein-DNA dissociation constant of protein  $i$  and enhancer region of gene  $i$ ;  $\mathbf{r}_i = (r_{i0}, r_{i1}, r_{i2})$  parameterizes the recruiting affinity of transcription factor  $A_{1,2}$  for the transcriptional machinery at genomic locus  $i$ ;  $\mathbf{t}_i = (t_{i1}, t_{i2})$  measures the relative strengths of the affinities of the transcription factors to enhancers of  $i$ : we set  $t_{ij} = (K_{di,A_j}/K_{d1,A_1})$ ,  $t_{i2} = (K_{di,A_2}/K_{d2,A_2})$ , for  $i = 1, 2, 3$ , so that  $t_{11} = 1 = t_{22}$ . The promoter occupancy probability  $\varphi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i)$  is defined in terms of the fold change function  $\psi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i)$  thus:

$$\begin{aligned} \varphi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i) &= \frac{1}{1 + \psi_i^{-1}(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i)} \\ \text{where } \psi_i(\mathbf{x}, \mathbf{r}_i, \mathbf{t}_i) &= \frac{r_{i0} + r_{i1}t_{i1}x_1^2 + r_{i2}t_{i2}x_2^2}{1 + t_{i1}x_1^2 + t_{i2}x_2^2} \end{aligned} \quad (\text{A.29})$$

Note, for simplicity we have used  $r$  and  $t$  instead of  $\tilde{r}$ ,  $\tilde{t}$  that we introduced to explicitly demonstrate the source of context dependence in the previous subsection.

## Appendix B. Parameter values for numerical experiments

The following parameters were used in the computations. Units of concentration in nM and time in hr.  $r_{a0} = r_{b0} = 1/1000$ ,  $\delta_a = 1 = \delta_b$ ,  $\Delta_B = 1/10$ ,  $\Delta_A = \Delta \times \Delta_B$ ,  $\mu_b = 2$ ,  $\mu_a = \mu_b\sqrt{c}$ ,  $\pi_b = 30$ ,  $\pi_a = \pi_b\sqrt{c}$ ,  $\kappa_A = \kappa_B = 5\sqrt{10}$ ,  $K_{da,A} = 1/2 = K_{db,B}$ ,  $\text{DA}_0 = 1 = \text{DB}_0$ .  $r_{ij}$  is taken to be  $10 \times r_0 \times r$  where the factor  $1 \leq r \leq 100$  is the ratio  $r_{ij}/r_{ik}$  which is taken as a variable, as are  $c$  ( $1 \leq c \leq 10$ ) and  $\Delta$  ( $1 \leq \Delta \leq 25$ ). For the switches, both heterozygous and homozygous, the ratio  $\Delta$  of degradation rates is taken to be 1 and so is  $c$  for the pre-duplication genotype.

## Appendix C. Hopf bifurcation surface via elimination

The fixed points of eq. (4),  $(x_1^*, x_2^*) = (c_1\Delta_2/c_2\Delta_1) = (c/\Delta)$  which are solutions to  $c_i\varphi - \Delta_i x_i = 0$ . Thus  $c_2\varphi(x_2, (c/\Delta)x_2) - \Delta_2 x_2 = 0$  is a cubic represented as  $f_1$  below. The eigenvalues of the Jacobian of the  $(x_1, x_2)$  system are evaluated at  $(x_1^*, x_2^*)$  to analyse stability; while negative eigenvalues



imply local stability, as above, analysis of instability for oscillatory solutions involves looking at complex eigenvalues  $\lambda = a \pm i|b|$  of the Jacobian of the dynamical system at its fixed points. At the fixed point, when the real part  $a$  goes from being negative (stable) to positive (unstable) as a continuous function of a parameter in the system, while  $|b| > 0$  the system is set to undergo a Hopf bifurcation and begins oscillating with a period  $2\pi/|b|$ . For a 2-dimensional system, the sum of the eigenvalues at the crossover value  $a = 0$  vanishes, making the trace of the Jacobian matrix 0, while the determinant remains positive in our case. The equations for the tracelessness of the Jacobian at the point  $x_1 = (c/\Delta)x_2$ :

$$J_{ij} = \frac{1}{(1 + \Psi)^2} (c_i \frac{\partial \Psi}{\partial x_j} - \Delta_i \delta_{ij})$$

where the partial derivatives are listed in eq. (11, reproduced below:

$$\begin{pmatrix} \partial/\partial\alpha_1 \\ \partial/\partial\alpha_2 \end{pmatrix} \Psi(\boldsymbol{\alpha}, \mathbf{t} = 1, \mathbf{r}) = \frac{(r_1 - r_2)}{(1 + \alpha_1 + \alpha_2)^2} \begin{pmatrix} \frac{r_1 - r_0}{(r_1 - r_2)} + \alpha_2 \\ \frac{r_2 - r_0}{(r_1 - r_2)} - \alpha_1 \end{pmatrix}. \quad (\text{C.1})$$

Thus the condition for a Hopf bifurcation reduces to solutions of

$$c_1 \frac{\partial \Psi}{\partial x_1} + c_2 \frac{\partial \Psi}{\partial x_2} = \Delta_1 + \Delta_2.$$

Using the simplification  $t_{ij} = 1$ , and upon substituting  $x_1 = (c/\Delta)x_2$ , this turns out to be a quartic equation  $f_2$  in  $x_2$  with coefficients being complicated, but polynomial combinations of the parameters. We compute the resultant of  $f_1, f_2$  to find a non-trivial greatest common divisor (gcd) and common root of these polynomials. The resultant is computed via the determinant of the Sylvester matrix of the polynomials  $f_1, f_2$ :

$$\begin{aligned} f_1 &= \sum_{i=0}^3 s_i x_2^i \quad \text{and} \quad f_2 = \sum_{i=0}^4 h_i x_2^i \\ s_0 &= \Delta^2 c_2 r_0, \\ s_1 &= -\Delta^2 (1 + r_0) \Delta_2, \\ s_2 &= (c^2 r + \Delta^2) c_2 r_2, \\ s_3 &= -(c^2 + \Delta^2 + c^2 r r_2 + \Delta^2 r_2) \Delta_2, \\ h_0 &= \Delta^4 (1 + \Delta) (1 + r_0)^2 \Delta_2, \\ h_1 &= 2\Delta^3 c_2 (r_0 + c^2 r_0 - r_2 - c^2 r r_2), \\ h_2 &= -2s_3 \Delta^2 (1 + \Delta) (1 + r_0), \\ h_3 &= -2c^2 (r - 1) \Delta (\Delta^2 - 1) c_2 r_2, \\ h_4 &= -s_3 (1 + \Delta) \end{aligned} \quad (\text{C.2})$$

given by

$$\begin{pmatrix} s_3 & 0 & 0 & 0 & h_4 & 0 & 0 \\ s_2 & s_3 & 0 & 0 & h_3 & h_4 & 0 \\ s_1 & s_2 & s_3 & 0 & h_2 & h_3 & h_4 \\ s_0 & s_1 & s_2 & s_3 & h_1 & h_2 & h_3 \\ 0 & s_0 & s_1 & s_2 & h_0 & h_1 & h_2 \\ 0 & 0 & s_0 & s_1 & 0 & h_0 & h_1 \\ 0 & 0 & 0 & s_0 & 0 & 0 & h_0 \end{pmatrix}.$$

This resultant factors into two pieces that change sign, and hence contains a zero. The first is

$$\left(c^2(r_1 - r_0) + \Delta^2(r_2 - r_0)\right)^2$$

which vanishes for

$$\frac{c}{\Delta} = \sqrt{-\frac{r_2 - r_0}{r_1 - r_0}}.$$

This condition implies that one of the two genes recruits Pol II more efficiently than the basal rate while the other's activation rate is less than that of basal transcription – *i.e.*, one is an activator, the other a repressor.

The second factor is a complicated function  $h(r, \Delta, c)$  and we omit it here. However, we can plot the surface  $h(r, \Delta, c) = 0$ , as shown in Figure 8. This is the case that we examine in depth in the paper.

## References

- [1] Ohno S. Evolution by Gene Duplication. Springer; 1970.
- [2] Lynch M. The Origins of Genome Architecture. Sinauer Associates Inc; 2007.
- [3] Clark AG. Invasion and maintenance of a gene duplication. Proceedings of the National Academy of Sciences. 1994;91(8):2950–2954.
- [4] Force A, Lynch M, Pickett FB, Amores A, Yan YI, Postlethwait J. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics. 1999;151(4):1531–1545.
- [5] Proulx SR, Phillips PC. Allelic divergence precedes and promotes gene duplication. Evolution. 2006;60(5):881–892.
- [6] Bourguet D. The evolution of dominance. Heredity. 1999;83:1–4.

- [7] Kondrashov FA, Koonin EV. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*. 2004;20(7):287–290.
- [8] Birchler JA, Yao H, Chudalayandi S, Vaiman D, Veitia RA. Heterosis. *Plant Cell*. 2010;22:2105–2112.
- [9] Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, et al. Core transcriptional regulatory circuitry in human hepatocytes. *Molecular Systems Biology*. 2006;2.
- [10] Davidson EH. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press; 2006.
- [11] Hobert O. Regulatory logic of neuronal diversity: Terminal selector genes and selector motifs. *Proceedings of the National Academy of Sciences*. 2008;105(51):20067–20071.
- [12] Johnston RJ, Desplan C. Stochastic Mechanisms of Cell Fate Specification that Yield Random or Robust Outcomes. *Annual Review of Cell and Developmental Biology*. 2010;26(1):689–719.
- [13] Tvrdik P, Capecchi MR. Reversal of Hox1 gene subfunctionalization in the mouse. *Dev Cell*. 2006;11:239–50.
- [14] Force A, Cresko WA, Pickett FB, Proulx SR, Amemiya C, Lynch M. The Origin of Subfunctions and Modular Gene Regulation. *Genetics*. 2005;170(1):433–446.
- [15] Wang Y, Schnegelsberg PNJ, Dausman J, Jaenisch R. Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin. *Nature*. 1996;379:823–5.
- [16] Kafri R, Levy M, Pilpel Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *PNAS*. 2006;103(31):11653–11658.
- [17] Otto SP, Yong P. 16 The evolution of gene duplicates. In: Dunlap JC, ting Wu C, editors. *Homology Effects*. vol. 46 of *Advances in Genetics*. Academic Press; 2002. p. 451 – 483.
- [18] Walsh B. Population-Genetic Models of the Fates of Duplicate Genes. *Genetica*. 2003;118:279–294. 10.1023/A:1024194802441.

- [19] Omholt SW, Plahte E, Uyehaug L, Xiang K. Gene Regulatory Networks Generating the Phenomena of Additivity, Dominance and Epistasis. *Genetics*. 2000;155(2):969–980.
- [20] Gjuvsland AB, Plahte E, Ådnøy T, Omholt SW. Allele Interaction – Single Locus Genetics Meets Regulatory Biology. *PLoS ONE*. 2010 02;5(2):e9379.
- [21] Spofford JB. Heterosis and the Evolution of Duplications. *The American Naturalist*. 1969;103(932):407–432.
- [22] Ni Z, Kim ED, Ha M, Lackey E, Liu J, Zhang Y, et al. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature*. 2009;457:327–331.
- [23] Dunlap JC. Molecular bases for circadian clocks. *Cell*. 1999;96:271–290.
- [24] Novak B, Tyson JJ. Design principles of biochemical oscillators. *Nat Rev Mol Cell Biol*. 2008 12;9:981–991.
- [25] Chen H, Xu L, Gu Z. Regulation Dynamics of WGD Genes during Yeast Metabolic Oscillation. *Molecular Biology and Evolution*. 2008;25(12):2513–6.
- [26] Trachana K, Jensen LJ, Bork P. Evolution and regulation of cellular periodic processes: a role for paralogues. *EMBO Rep*. 2010;11(3):233–238.
- [27] Baggs JE, Price TS, DiTacchio L, Panda S, FitzGerald GA, Hogenesch JB. Network Features of the Mammalian Circadian Clock. *PLoS Biol*. 2009 03;7(3):e1000052.
- [28] Shi S, Hida A, McGuinness OP, Wasserman DH, Yamazaki S, Johnson CH. Circadian Clock Gene *Bmal1* Is Not Essential; Functional Replacement with its Paralog, *Bmal2*. *Current biology*. 2010;20(4):316–321.
- [29] Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences of the United States of America*. 1982;79(4):1129–1133.
- [30] Cherry JL, Adler FR. How to make a biological switch. *Journal of Theoretical Biology*. 2000;203:117–133.
- [31] Phillips RB, Kondev J, Theriot J. *Physical biology of the cell*. Garland Science; 2009.

- [32] Ptashne M, Gann A. Genes and Signals. Cold Spring Harbor Laboratory Press; 2001.
- [33] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*. 2005;15(2):116 – 124.
- [34] Bergthorsson U, Andersson DI, Roth JR. Ohno’s dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences*. 2007;104(43):17004–17009.
- [35] Raj A, van Oudenaarden A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*. 2008 October;Volume 135(2):216–226.
- [36] Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature*. 2010;467(7312):167–173.
- [37] Wang Z, Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences*. 2011;108(16):E67-E76.
- [38] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genetical Research*. 1966;8:269–94.
- [39] Hartl DL, Clark AG. *Principles of Population Genetics*, Fourth Edition. Sinauer Associates Inc; 2007.
- [40] Bundschuh R, Hayot F, Jayaprakash C. Fluctuations and Slow Variables in Genetic Networks. *Biophys J*. 2003;84:1606–1615.
- [41] Del Vecchio D, Ninfa AJ, Sontag ED. Modular cell biology: retroactivity and insulation. *Mol Syst Biol*. 2008;4(10.1038/msb4100204).
- [42] Pakka VH. Dynamics of molecular fluctuations in gene regulatory networks. University of Southampton, UK; 2009.
- [43] Dasmahapatra S. Oscillatory dynamics in a double activator motif. In: *International Workshop on Computational Systems Biology* 2011.
- [44] Guantes R, Poyatos JF. Multistable Decision Switches for Flexible Control of Epigenetic Differentiation. *PLoS Comput Biol*. 2008 11;4(11):e1000235.

- [45] Strogatz SH. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Westview Press; 2000.
- [46] Stoltzfus A. On the Possibility of Constructive Neutral Evolution. *Journal of Molecular Evolution*. 1999;49(2):169–81.
- [47] Haldar M, Karan G, Tvrdik P, Capecchi MR. Two Cell Lineages, myf5 and myf5-Independent, Participate in Mouse Skeletal Myogenesis. *Developmental Cell*. 2008 March;14(3):437–445.
- [48] Wang X, Muratani M, Tansey WP, Ptashne M. Proteolytic Instability and the Action of Nonclassical Transcriptional Activators. *Current Biology*. 2010;20(9):868 – 871.
- [49] Armstrong RA, McGehee R. Competitive Exclusion. *The American Naturalist*. 1980;115(2):151–170.
- [50] Cox D, Little J, O’Shea D. *Ideals, Varieties, and Algorithms*. Springer; 2007.
- [51] Hasty J, Dolnik M, Rottschäfer V, Collins JJ. Synthetic Gene Network for Entraining and Amplifying Cellular Oscillations. *Phys Rev Lett*. 2002 Mar;88(14):148101.
- [52] Stricker J, Cookson S, Bennett M, Mather W, Tsimring L, Hasty J. A fast, robust and tunable synthetic gene oscillator. *Nature*. 2008;456:516–519.
- [53] Winfree A. *The Geometry of Biological Time*. Springer; 2001.
- [54] Hong CI, Conrad ED, Tyson JJ. A proposal for robust temperature compensation of circadian rhythms. *Proceedings of the National Academy of Sciences*. 2007;104(4):1195–1200.
- [55] Gillespie DT. A General Method for Numerically Simulating the Stochastic Time evolution of Coupled Chemical Reactions. *Journal of Computational Physics*. 1976;22:403–434.
- [56] Adams KL, Wendel JF. Allele-Specific, Bidirectional Silencing of an Alcohol Dehydrogenase Gene in Different Organs of Interspecific Diploid Cotton Hybrids. *Genetics*. December 2005;171(4):2139–2142.
- [57] Tsuchiya M, Ross J. Advantages of external periodic events to the evolution of biochemical oscillatory reactions. *Proceedings of the National Academy of Sciences*. 2003;100(17):9691–9695.

- [58] Reijenga KA, Snoep JL, Diderich JA, van Verseveld HW, Westerhoff HV, Teusink B. Control of glycolytic dynamics by hexose transport in *Saccharomyces cerevisiae*. *Biophysical Journal*. 2001;80(2):626–634.
- [59] Brown CJ, Todd KM, Rosenzweig RF. Multiple Duplications of Yeast Hexose Transport Genes in Response to Selection in a Glucose-Limited Environment. *Mol Biol Evol*. 1998;15(8):931–942.
- [60] Rutter J, Reick M, McKnight SL. Metabolism and the control of circadian rhythms. *Annual Review of Biochemistry*. 2002;71(1):307–331.
- [61] Dodd AN, Salathia N, Hall A, Kevei E, Toth R, Nagy F, et al. Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science*. 2005;309(5734):630–633.
- [62] Ukai-Tadenuma M, Kasukawa T, Ueda HR. Proof-by-synthesis of the transcriptional logic of mammalian circadian clocks. *Nat Cell Biol*. 2008;10(10):1154–63.
- [63] Muratani M, Tansey WP. How the ubiquitin-proteasome system controls transcription. *Nat Rev Mol Cell Biol*. 2003;4(3):192–201.
- [64] Munteanu A, Constante M, Isalan M, Sole R. Avoiding transcription factor competition at promoter level increases the chances of obtaining oscillation. *BMC Systems Biology*. 2010;4(1):66.
- [65] Tu B, Kudlicki A, Rowicka M, McKnight S. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*. 2005;310:1152–8.

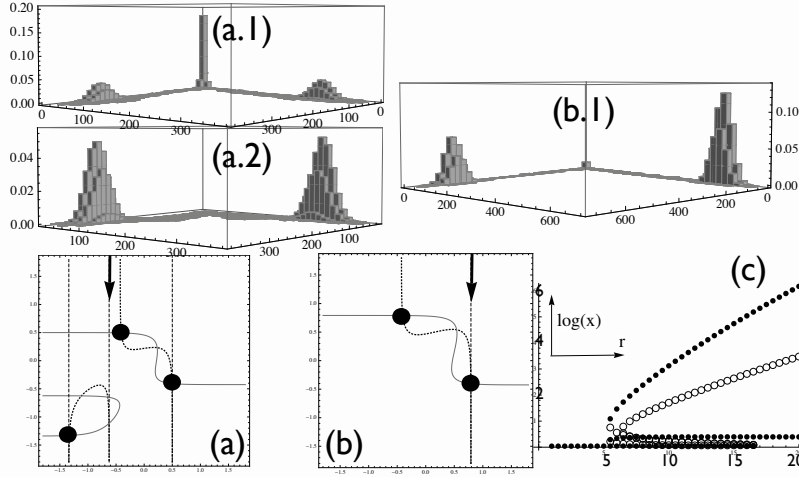
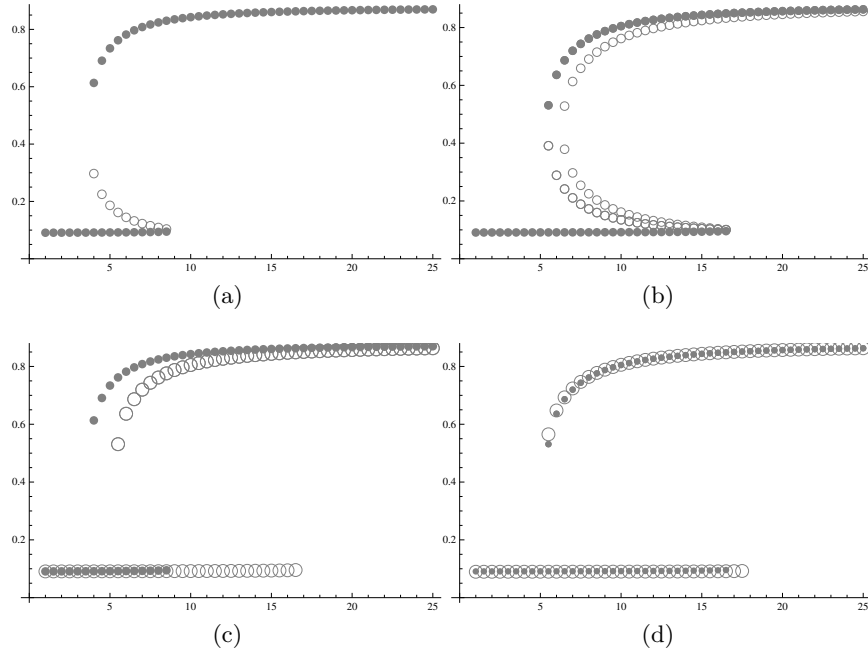


Figure 6: **The (heterozygous) context dependent dual-activator switches between three different expression states.** The  $x_1$ - $x_2$  planes in this figure are in logarithmic scale. For the case  $r_{11} = r_{22} = r_1$ ,  $r_{21} = r_{12} = r_2$ , we obtain multiple steady states. Here we illustrate the transition between three and two steady states as the ratio  $r = r_1/r_2$  changes from a value of 10 in (a.1) and (a.2) to 20 in (b.1). All other parameter values are as in Appendix B, with  $c = \Delta = 1$ , and the activation strengths are multiples,  $r$  of  $r_0$ . The two histograms in (a.1) and (a.2) correspond to different initial conditions on either side of the arrow drawn in the phase-plane (a) for small noise (large  $\Omega$ ). The vertical lines in (a) and (b) correspond to the case of the single gene switch. The outward point double-arrow in (a) shows the threshold for the single gene case (which is close to that of the duplicated gene as well), while the inward pointing single arrow locates the only (stable) fixed point for large  $r$  in the single gene case. On the right are shown the stable and unstable fixed points of  $x_1, x_2$  (ordinate) for different values of  $r$  (abscissa). Because of symmetry  $x_1$  and  $x_2$  share the same steady state values, but not concurrently.





**Figure 7: (Expression levels vs  $r$ ). The decision switch is indistinguishable from the single gene switch in downstream effects.** In each of these figures, the abscissa indicates the value of  $r$  in the two-gene subnetwork. In (a) we have the case of unchanged duplicated autoregulatory gene  $r_{ij} = r$ , while in (b) we have the case  $r_{11} = r_{22} = r$  while  $r_{12} = 10r_0 = r_{21}$  where the basal rates are taken to be the same  $r_{i0} = r_0$ . The open circles in the top two figures (a) and (b) are the values of the target gene  $z$  corresponding to the unstable fixed points. The stable fixed point values are in filled circles. The stable values of  $z$  in (a) and (b) are gathered together in (c) for comparison, with the open circles correspond to values in (b), filled circles to those in (a). Finally, in (d) we compare the stable values from the context-dependent case (b) to those from the target expression of the pre-duplicated system. Note that if we use the expression of the target gene as a readout of the effect of gene duplication, we find no change (see (d)) prior to further divergence. For the decision switch, (c) shows that the only qualitative change lies in the threshold for switching and hysteresis.

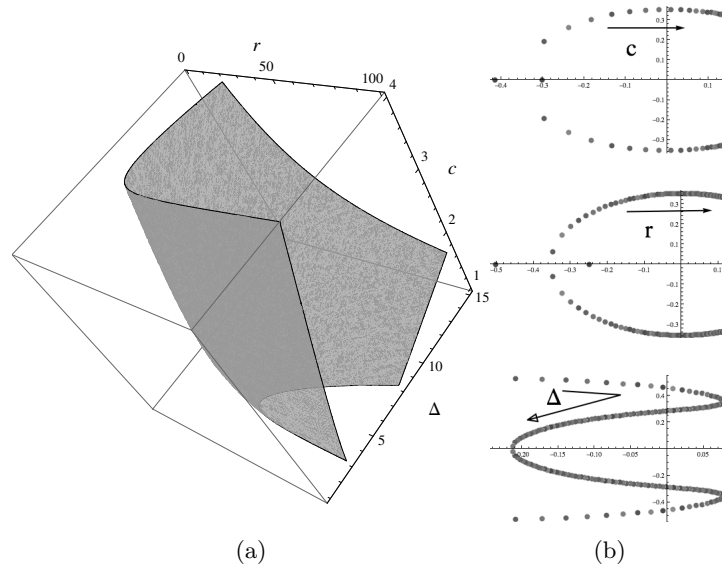


Figure 8: **Parameter space where Hopf bifurcation occurs.** As explained in the text, eliminating  $x_2$  from  $f_1$  and  $f_2$  for a set of parameter values  $c_2 = 50$ ,  $r_0 = 10^{-3}$  and  $r_2 = 10^{-2}$  yields a surface in  $(r, \Delta, c)$  3-dimensional parameter space depicted in (a). The rest of the parameters are as in Appendix B. In (b) we show the complex plane of the eigenvalues with two of the three parameters fixed, and the increase of the third through the bifurcation point(s). This occurs as  $c, r$  increases, and the non-monotonic increase of the positive real part of the eigenvalue as  $\Delta$  increases is also clearly seen at the bottom of (a).



Figure 9: The two topologies obtained by solving for the Hopf bifurcation condition. The one on the left is the subnetwork we address in this paper.

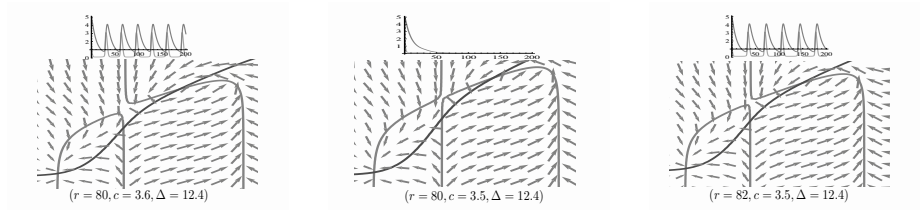


Figure 10: **Global bifurcations** leading to oscillations are exemplified in this figure, where an excitatory trajectory can be trapped into an oscillatory state by realignment of nullclines upon parameter changes. This can happen for both, changes of  $r$  and of  $c$ : the middle panel has parameter values  $\theta^* = (r = 80, c = 3.5, \Delta = 12.4)$ , whereas the left panel has  $\theta + \delta c$  and the right panel has  $\theta + \delta r$ , with  $\delta c = 0.1$  and  $\delta r = 2$ . Either change modifies the  $x_1$  nullcline from  $\cup \rightarrow \cap$ . Time series plots for each case is shown above the phase planes.

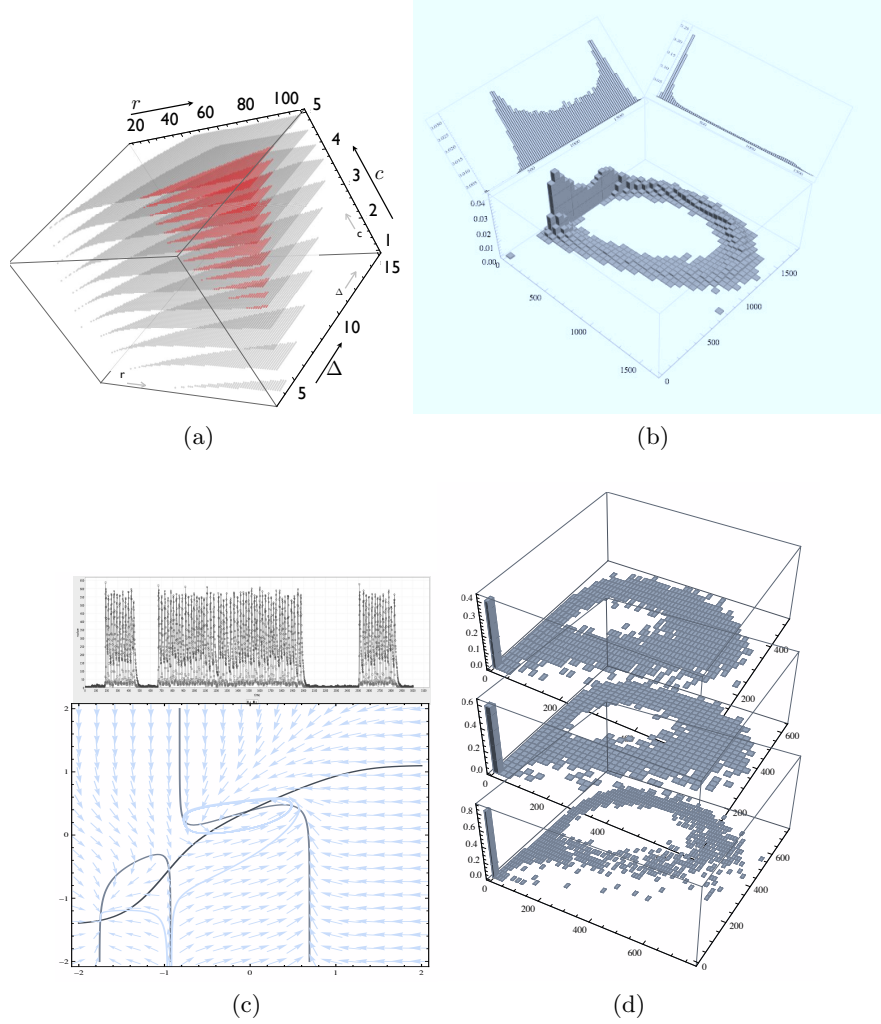


Figure 11: **Bursts of oscillations are controlled by the amount of noise in expression.** The  $x_1$ - $x_2$  planes in this figure are in logarithmic scale. (a): Oscillatory regimes in parameter space – red (dark online) region has a stable limit cycle; the light (grey online) has a coexisting steady state as well. (b): Histograms accumulated from a simulation of the stochastic kinetics in the red oscillatory phase with  $(x_1, x_2, P(x_1, x_2))$  as the right-handed coordinate axes. Also shown are histograms for  $x_1$  (unimodal) and  $x_2$  (bimodal). (c): Phase plane with  $x_1 - x_2$  nullclines and the trajectories of two different initial points, chosen to be on either side of the separatrix determined by the saddle-node (in the middle of the three intersections). (d): Histograms of trajectories in the light grey region of (a) for different noise strengths (which decreases from top to bottom) showing the redistribution of probability mass around the limit cycle away from low-expression peak.