Error-Minimizing As Nature's Underlying Design Principle For Gene Regulation

Jialing Li

Department of Physics, Massachusetts Institute of Technology
Email: jialing3@mit.edu

(Submitted to 8.592 as a final project, May 2007)

Transcriptional control of gene expression can be either positive or negative, achieved separately through an activator or a repressor binding to the regulatory site. Shinar *et al.* proposed that regardless of the mode of control, evolution favors keeping the gene regulatory sites bound for most of the time, since the open sites are prone to non-specific binding errors[1]. Their model incorporates error-related fitness advantage difference of the two regulation modes into the Savageau demand rule and demonstartes that error-minimizing could drive the evolution of transcriptional control. The mathematical model was futher tested for the *E. coli lac* system and shown to be applicable to multi-input gene regulatory systems.

I. Introduction

Like any regulatory systems, gene expression involves converting various input signals into specified output signal. How do biological systems achieve transcriptional control and keep high-fidelity of their gene product in a noisy cellular environment? Cells use regulatory proteins that bind specific sites in the promoter region of the regulated genes. Transcription of a gene is fully turned "on" by two mechanisms: the positive control needs an activator to bind and the negative control requires an repressor to unbind. A gene can have multiple regulators, each of which can be either an activator or a repressor, to receive multiple inputs. There are 2^N possible mechanisms in total for a gene controlled by N regulators. In the second section of this paper, we will study a general single regulator system and one specific multi-regulator system in *E. coli*.

Are all the regulation mechanisms equivalently favorable in nature? It turns out, as the Savageau rule states for *E. coli* as well as other organisms[2-5], high-demand genes tend to have positive control whereas low-demand genes is more negatively regulated. Particularly, in well-characterized catabolic pathways that break down sugar sources, positive control is more abundant for sugars commonly found in the natural environment, and negative control more for rarely seen sugars[3,5,16].

Savageau demand rule also states that negative(positive) control will be selected against for a high(low)-demand gene[3,5]. For instance, if a gene is needed constantly, mutations eliminating the repressor is hardly selected against because their expression is kept "on" and thus its cellular function is not disrupted; as opposed to mutations that get rid of the activator is deleterious and readily selected against due to their low expression level. For a low-demand gene, the opposite is true. These assumptions remain valid as long as there is no inherent fitness advantage to one mode of control.

However, operating in a noisy cellular environment, the two modes of control are indeed subject to inherent fitness differences. Whenever a regulatory site is free (not bound by its cognate regulator), it is error-prone in a sense that leaky transcription can occur by unspecific binding of other transcription factors. The errors in gene-expression levels will then lower the whole organism's fitness. *Shinar et al.* claims that in order for a biological system to minimize the fitness-reducing error, it should keep the

regulatory sites bound by cognate regulators for most of the time[1]. Such a system tend to evolve positive control in high-demand environments and negative control in low-demand environments.

The relative reduction of fitness due to non-specific bindings and leaky transcriptions is called the *error-load*. There are three major sources associated with the non-specific bindings: cross-talk with other transcriptional regulators[6-9], lateral gene transfer[10,11], and residual binding of the designated regulator in an inactive form. Generally, cellular environment differences can lead to large variations in residual binding and thus give rise to cell-to-cell fluctuations in gene expression.

II. A Mathematical Model and Its Application To A Multi-Regulator System in *E. coli*

It is easy to compare the error-load of the positive and negative modes of control for the regulated gene with a demand p. p is the fraction of time the full expression of the gene product is needed in the environment. Hence, the fraction of time of a regulatory site being exposed is p for the repressor and $(1-p)$ for the activator. Suppose the relative reduction of fitness is Δf_1 for the high expression state and Δf_0 for the low expression state, then the average error-loads, taking into account the errors from the free sites only, can be expressed as the following.

$$E_R = p\,\Delta f_1 \qquad [1]$$

$$E_A = (1-p)\,\Delta f_0 \qquad [2]$$

In order for repressor to have a lower error-load, we demand $E_R < E_A$. This leads to,

$$p < p^* = 1/(1+\Delta f_1/\Delta f_0) \qquad [3]$$

Equation [3] translates into the demand must remain below a certain threshold p^* for the repressor mode to be more advantageous. Therefore, low-demand genes favor negative(repressor) regulation, whereas high-demand genes favor positive(activator) regulation. See figure 1 for a plot of demand threshold versus $\Delta f_1/\Delta f_0$.

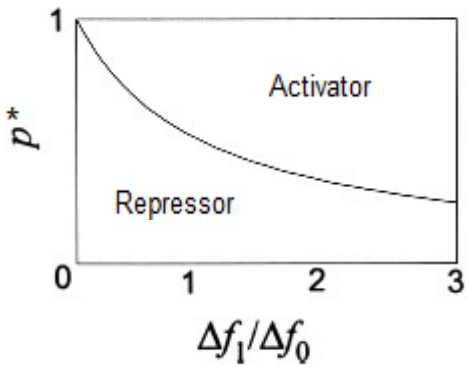

Fig. 1

Since there is a difference between the error-loads of the positive and negative controls, shifts in environmental conditions can then exert selective pressure on an organism for it to optimize its gene regulatory system and to minimize the error-load. Mutations can get fixed if their relative fitness advantage $|E_A - E_R|$ exceeds a minimal selection threshold S_{min}, which is estimated to be 10^{-8} to 10^{-7} in bacteria[12,13]. For $S_{min} < E_A - E_R$, a repressor mutant gets fixed; for $S_{min} < E_R - E_A$, an activator mutant gets fixed. Plug equations [1] and [2] into these conditions, we can see that for $S_{min} < (\Delta f_0 + \Delta f_1)$, the error-minimizing mode of control will be selected for. However, for $S_{min} > (\Delta f_0 + \Delta f_1)$, no mutations can be fixed, i.e. the historical precedent rules. Experimental data show that in yeast and *E. coli* 1-2% error in transcription can lead to Δf_0 and Δf_1 large enough to cause the fitness difference to exceed the selection threshold[12,14]. So, even small amount of transcriptional error can give rise to selectable error-load differences.

We now arrive at a revised version of the demand rule with errors caused by leaky transcription taken into consideration for a single regulator. Minimization of error serves as a driving force for high-demand genes to get positively controlled and low-demand genes to get negatively regulated. This rule can be extended to multiple-regulator systems as well[1]. Consider the *E. coli lac* system with multiple input signals for example. The cells express Lac

transporter protein in the presence of lactose alone. Lac transporter expression is inhibited in the presence of glucose. The Lac transporter gene has a repressor LacI and an activator CRP. The repressor action is modified by lactose and the activator is modified by glucose. CRP bound by glucose can no longer bind to the regulatory site, so the expression level of Lac transporter decreases when glucose is available. LacI bound by lactose can no longer inhibit transcription, so the cells start expressing Lac transporters when lactose becomes present. When both sugars are present, the cells have additional mechanisms to block the entry of lactose to ensure the consumption of glucose[15].

There are four possible states of the combined binding of LacI and CRP. [CRP, LacI] = [0,0], [0,1], [1,0], [1,1], where 1 is bound and 0 is unbound. Since only one sugar can be present inside the cells, the [0,0] state is not achievable. The wild-type *lac* system has a lactose-responsive repressor and a glucose-responsive activator. So, we have [glucose-responsive regulator mode, lactose-responsive regulator mode] = [A,R] for the wild-type, where A stands for activator and R for repressor. In theory, there exist three other modes of control as [R,R], [R,A], [A,A]. Each of these four possible modes of control has one excluded binding state. Recall for the wild type [AR], the [00] state is unachievable. So is [01] for [AA], [10] for [RR], and [11] for [RA]. The fitness reductions due to one or two regulatory sites being free for different modes of control under different input conditions are listed in table 1[1].

Mechanism/input state	(0,0)	(0,1)	(1,0)/(1,1)
AA	$\Delta f_2'$	0	$\Delta f_1 + \Delta f_1'$
AR	0	$\Delta f_4'$	Δf_1
RA	$\Delta f_2 + \Delta f_2'$	Δf_4	$\Delta f_1'$
RR	Δf_2	$\Delta f_4 + \Delta f_4'$	0

Table 1

The rows are the four mechanisms of regulation, and the columns correspond to the input states (glucose, lactose). The unprimed relative reduction in fitnesses denote those of the glucose-responsive regulators and the primed ones are for lactose, each with index i = 1, 2, 3, 4 denoting the level of expressions under different input conditions. In particular, for the wide-type AR mode with input of (0,1), the binding states [CRP, LacI] will be [1,0]. This input state have the highest expression level, thus i = 4 for it. And since only the lactose-responsive regulator site is unbound, the relative reduction in fitness is $\Delta f_4'$.

The next step is to calculate the individual error-loads for all regulatory mechanisms. This is done by summing the average fitness reductions over all input sates.

$$E_{AR} = p_{01}\Delta f_4' + (1- p_{00} -p_{01})\Delta f_1 \quad [4]$$

$$E_{AA} = p_{00}\Delta f_2' + (1- p_{00} -p_{01})(\Delta f_1 + \Delta f_1') \quad [5]$$

$$E_{RA} = p_{00}(\Delta f_2 + \Delta f_2') + p_{01}\Delta f_4' + (1- p_{00} -p_{01})\Delta f_1' \quad [6]$$

$$E_{RR} = p_{00}\Delta f_2 + p_{01}(\Delta f_4 + \Delta f_4') \quad [7]$$

where p_{00} is the fraction of time that no sugar is present in the cells and p_{01} is the fraction of time that only lactose is present.

The same method used to compare error-loads of diffenrent modes of a single regulator can be adapted to compare the 4 modes of this multi-regulator system under a given cellular environment (p_{00}, p_{01}). The results are summarized graphically in figure 2. Basially, the RA mechanism never has the lowest error-load and so is not favored. On the other hand, the wile-type, AR mechanism, is capable of minimizing the error-load in the absence of both sugars ($p_{00} \sim 0$, $p_{01} \ll 1$). Since indeed both

glucose and lactose are not abundant in *E. coli*'s natural environment[2,5], the glucose-responsive activator and lactose-responsive repressor mechanism is the fittest in terms of minimizing error-load resulting from non-specific binding.

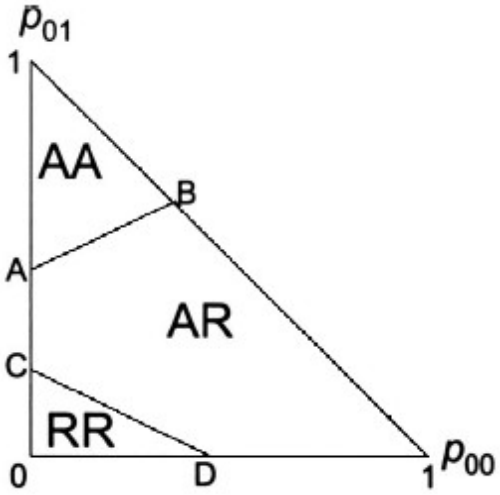

Fig. 2 This is a triangle because $p_{00} + p_{01} \leq 1$. Line AB has a slope of $(\Delta f_2' - \Delta f_1')/(\Delta f_1' + \Delta f_4')$ and intercepts the vertical axis at $\Delta f_1'/(\Delta f_1' + \Delta f_4')$. Line CD has a slope of $-(\Delta f_1 + \Delta f_2)/(\Delta f_1 + \Delta f_4)$ and intercepts the vertical axis at $\Delta f_1/(\Delta f_1 + \Delta f_4)$.

The multi-regulator picture might seem more complicated than the single-regulator case, but the underlying principles are the same. The binding of cognate regulators reduces error, because free binding sites are subject to non-specific bindings. As for a predominant input of (0,0), i.e. no glucose or lactose, the AR mechanism ensures the binding of both regulatory sites, whereas AA and RR will expose one site. Because the RA mechanism exposes both binding sites under the (0,0) input, it will never minimize the error-load. Moreover, the cells will never have both sugars simultaneously present, and this allows the AR mechanism to at most have one free binding site, reducing its overall error-load.[1]

III. Generalization and Future Experimental Test of The Model

Shinar *et al.*'s theory of error-minimization being an underlying principle for biological regulation can be generalized to other systems that requires sequence-specific binding among biomolecules, for instance, protein-protein interactions[1].

Furthermore, this theory provides strong motivation to an experimental test, to search for a potential design principle for gene circuits. One can construct synthetic systems of desired regulation mechanisms in yeasts or *E. coli* and allow them to either compete or to evolve under pressure of changing environment.

Acknowledgement

This work owes much to course 8.592 taught by Professor Kardar and Professor Mirny at MIT.

Reference

1. G. Shinar, *et al.*, PNAS **103**, 3999 (2006).
2. M. A. Savageau, PNAS **71,** 2453 (1974).
3. M. A. Savageau, PNAS **74,** 5647 (1977).
4. M. A. Savageau, PNAS **80,** 1411 (1983).
5. M. A. Savageau, Genetics **149,** 1665 (1998).
6. R. S. Rabin and V. Steward, J. Bact **175,** 3259 (1993).
7. M. T. Martinez-Pastor, *et al.,* EMBO J. **15,** 2227 (1996).
8. A. M. Sengupta, M. Djordjevic and B. I. Shraiman, PNAS **99,** 2072 (2002).
9. N. E. Buchler, U. gerland and T. Hwa, PNAS **100,** 5136 (2003).
10. H. Ochman, J. G. Lawrence and E. A. Groisman, Nature **405,** 299 (2000).
11. H. H. McAdams, B. Srinivasan and A. P. Arkin, Nat. Rev. Genet. **5,** 169 (2004).
12. A. Wagner, Mol. Biol. Evol. **22,** 1365 (2005).
13. D. L. Hartl, E. N. Moriyama and S. A. Sawyer, Genetics **138,** 227 (1994).
14. E. Dekel and U. Alon, Nature **436,** 588 (2005).
15. P. W. Postma, J. O. Lengeler and G. R. Jacobson, Microbiol. Rev. **57**, 543 (1993).
16. M. A. Savageau, Chaos **11**, 142 (2001).