bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transcriptional crosstalk varies between regulatory strategies

Rok Grah(1) and Tamar Friedlander(1,2)

(1) Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria

(2) The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture

Faculty of Agriculture, Hebrew University of Jerusalem,

P.O. Box 12 Rehovot 7610001, Israel

July 19, 2018

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Biological functions can often be realized by multiple network architectures, that require different numbers of molecular species. Here, we focus on the example of negative vs. positive gene regulation. While these distinct architectures can both implement the same input-output relation, they incur different component usage. We analyze how the choice of either form of regulation affects global transcriptional crosstalk, when genes are differentially expressed. Previous works proposed that the form of regulation correlates with gene usage: positive if the gene is frequently expressed and negative when infrequently expressed. This implies that in the absence of regulation genes are typically found in their unrequired activity state, hence regulatory intervention is often necessary. We find that such excess use of regulation usually entails high crosstalk levels. Instead, we propose that crosstalk reduction could be facilitated by the opposite parsimonious strategy, where genes are in their frequently required activity state, when not employing any form of regulation. Our study demonstrates the pervasive impact of a new type of protein production cost which is typically overlooked: that of regulatory interference caused by the presence of excess DNA-binding proteins in the cellular medium. This regulatory cost is predicted to increase with the number of molecular species and interactions involved. Hence, it only becomes apparent, when multiple transcription factors and multiple genes are simultaneously considered.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author Summary

Gene regulation can take two basic forms: positive - namely, the gene is inactive unless a regulator activates it; or negative - such that it is inactive unless a regulator suppresses its activity. A long-standing question is why some genes are positively, while others are negatively regulated. Savageau's demand rule proposed, that the form of regulation correlates with the gene's usage: positive if it is frequently active and negative when infrequently active. This implies, however, that in the absence of regulation genes are typically found in their least required activity state, hence regulatory intervention is often needed. In this work, we studied how the form of regulation - positive or negative - affects crosstalk levels, using a mathematical model for global crosstalk in a many-genes many regulators setup. By crosstalk, we mean that genes are occasionally found in an undesired regulatory state, because they interact with one of the many foreign regulators available. We find that crosstalk usually increases with the availability of regulators. Crosstalk can be reduced by the opposite parsimonious strategy, where genes are in their frequently required activity state, when not employing any form of regulation. This study demonstrates that crosstalk is a new type of global constraint on gene regulatory networks.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Gene regulatory networks can employ different architectures that seemingly realize the same input-output relation. There is a basic dichotomy of gene regulation into positive and negative control. A gene controlled by positive regulation is, by default, not expressed and requires binding of an activator to its operator to induce it. In contrast, a gene controlled by negative regulation, is expressed by default, unless a repressor binds its operator and attenuates its activity. While a gene can be activated or inactivated with either mode of regulation, researchers have pondered whether additional considerations could favor the choice of one mechanism over the other, or whether this choice is merely a coincidence ("evolutionary accident"). Throughout the years, this question received attention from different researchers, employing different approaches. It was first addressed in the seminal work of Michael Savageau [1, 2, 3], who proposed the so-called "Savageau demand rule": namely, that genes encoding frequently needed products ("high-demand") are often regulated by activators; conversely, genes whose products are only needed sporadically ("low-demand") tend to be regulated by repressors. This finding was based on a survey of a limited set of bacterial genes known at that time. Savageau used an evolutionary argument to explain this finding, later called the "use-it-or lose-it" principle [4]. He considered mutations and selection affecting the regulatory construct, and argued that the intensity of selection depends on the extent to which it is used. If the regulatory construct is infrequently used (as in activator regulating a low-demand or a repressor regulating a high-demand gene), selection on it is weak and it is unlikely to survive. In contrast, if it is frequently used (activator regulating a high-demand or repressor regulating a low-demand gene), selection is stronger and this form of regulation is more likely to survive [5]. A later evolutionary analysis mathematically formulated the problem as selection in an alternating environment and found the exact conditions under which the Savageau demand rule is expected to hold, in terms of population size and time between environmental switches in which the gene product is demanded or not [4].

An alternative reasoning for the observed correlation between a gene's demand and its form of regulation was later proposed using a biophysical, rather than evolutionary, argu-

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ment [6, 7]. There, it was suggested, that the Savageau demand rule minimizes transcriptional crosstalk, because the regulatory binding sites are mostly occupied by their cognate regulators. When occupied, these sites are protected from random binding of foreign proteins that could interfere with the gene's regulatory state. A possible critique of this reasoning is the extravagant use of regulators it entails: both high- and low-demand genes require regulators most of the time, which would place heavy demands on protein expression systems, associated with reduced organismal fitness [8, 9, 10].

Recently, a comprehensive survey of regulatory topologies in *E. coli* and *B. subtilis* found many exceptions to the Savageau demand rule. The authors proposed that a regulatory topology is simply randomly picked from a pool of networks meeting the physiological constraints [11].

While the above mentioned studies examined the significance of regulatory architectures from different perspectives, they all concentrated on a single gene with a single regulator (with potential external interference). They did not consider the additional regulatory interactions between each regulator and all other genes at play. While each interaction alone could be weak, an enormous number of such interactions occur and their total effect is significant [12, 13]. Here we focus on the effects of transcriptional crosstalk and how it might dictate the choice between different regulatory topologies. We specifically refer to *global crosstalk*, which accounts for crosstalk in all genes simultaneously, rather than the single-gene crosstalk considered earlier.

We use a mathematical model for global crosstalk [13], where we build upon the well-established thermodynamic model of gene regulation to calculate transcription factor (TF)-DNA interactions [14, 15, 16, 17, 18, 19]. We have previously shown that while crosstalk affecting a particular gene can be reduced by different means, it always comes at the cost of elevating crosstalk in other genes; the global crosstalk cannot be reduced below a certain threshold. We analyze global crosstalk levels under different regulatory strategies: either positive or negative regulation. We compare two extreme strategies: a "busy" one that implements the Savageau demand rule, in which a high (low)-demand gene is always regulated by an activator (repressor) and an opposite "idle" strategy, in which a high (low)-demand gene is always regulated by a repressor (activator). We find that the "busy" strategy maxi-

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mizes regulator usage, whereas the "idle" one minimizes it. Since global crosstalk depends on the abundance of regulatory proteins in the cellular environment, we conclude that under most biologically plausible parameter values, the "idle", but not the "busy" strategy should yield lower *global* transcriptional crosstalk.

This paper begins with the introduction of a general symmetric model for the analysis of transcriptional crosstalk in a many-TFs-many-genes setting, with combination of positive and negative regulation. We show that global crosstalk levels directly depend on the fraction of TFs in use and only indirectly on the choice of activation or repression as the form of regulation. We then analyze TF usage and crosstalk levels of the two extreme strategies, i.e., "busy" and "idle". We append with numerical simulations of a more general asymmetric gene usage model, that are in agreement with the analytical result. In the Box, we discuss the challenges in crosstalk calculation for real gene regulatory networks and show an example using data from *S. cerevisiae*.

Results

A model of gene regulation using a combination of activators and repressors

We consider a cell that has a total of M genes, each of which is transcriptionally regulated to be either active or inactive. We assume that each gene is regulated by a single unique TF species - its cognate one. Each gene has a short DNA binding site to which its cognate TF binds. A fraction $0 \leqslant p \leqslant 1$ of the genes is regulated by activators and the remaining $1-p$ fraction of genes is regulated by repressors. When no activator is bound, activator-regulated genes are inactive (or active at a low basal level) and only become active once an activator TF binds their binding site. In contrast, repressor-regulated genes are active, unless a repressor TF binds their binding site and inhibits their activity (Fig 1A). We assume that different environmental conditions require the activity of different subsets of proportion $0 \leqslant q \leqslant 1$ of these genes, while the remaining $1-q$ fraction remains inactive. These activity states are regulated by the binding and unbinding of the TFs specialized for these genes. We assume that only a subset of TFs necessary to maintain the desired regulatory pattern,

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are available to bind and regulate these genes. TFs however often have limited specificity to their DNA targets and can occasionally bind slightly different sequences, albeit with lower probability [20, 21, 22, 23, 24].

We define 'crosstalk', which potentially leads to an undesired regulatory outcome, as cases in which a binding site that should be bound by a specific TF is instead bound by a non-cognate one or remains unbound (x_{bound}), or in which a binding site that should have been unbound is occupied (x_{unbound}) - see Fig 1C. To quantitate the probability of these events, we use the thermodynamic model of gene regulation [14, 15, 16, 17]. A mathematical model for crosstalk for the special case in which all TFs are activators ($p = 0$) was derived and analyzed in a previous work [13]. Here, we analyze a more general model with a combination of activators and repressors. The reader can find the details of both models in the SI of this paper.

Both activity and inactivity of genes can be attained by means of either activator or repressor regulation. Accordingly, our model distinguishes between four sets of genes (see Table below and Fig 1B):

activity	regulated by	proportion of genes using this regulatory strategy
active	activator	a, where $a \leqslant q, p$
active	repressor	$q - a$
inactive	activator	$p - a$
inactive	repressor	$(1 - p) - q + a$

The probability that a particular gene i is in the x_{bound} or x_{unbound} crosstalk states, depends on the copy number of competing non-cognate TFs, C_j, $j \neq i$ and on the number of mismatches, d_{ij} between each competing TF j and the regulatory binding site of gene i, where we assume equal energetic contributions of all positions in the binding site. Consequently, the similarity between binding sites regulated by distinct TFs is a major determinant of crosstalk. We introduce an average measure of similarity between binding site i and all other binding sites $j \neq i$ [13]:

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

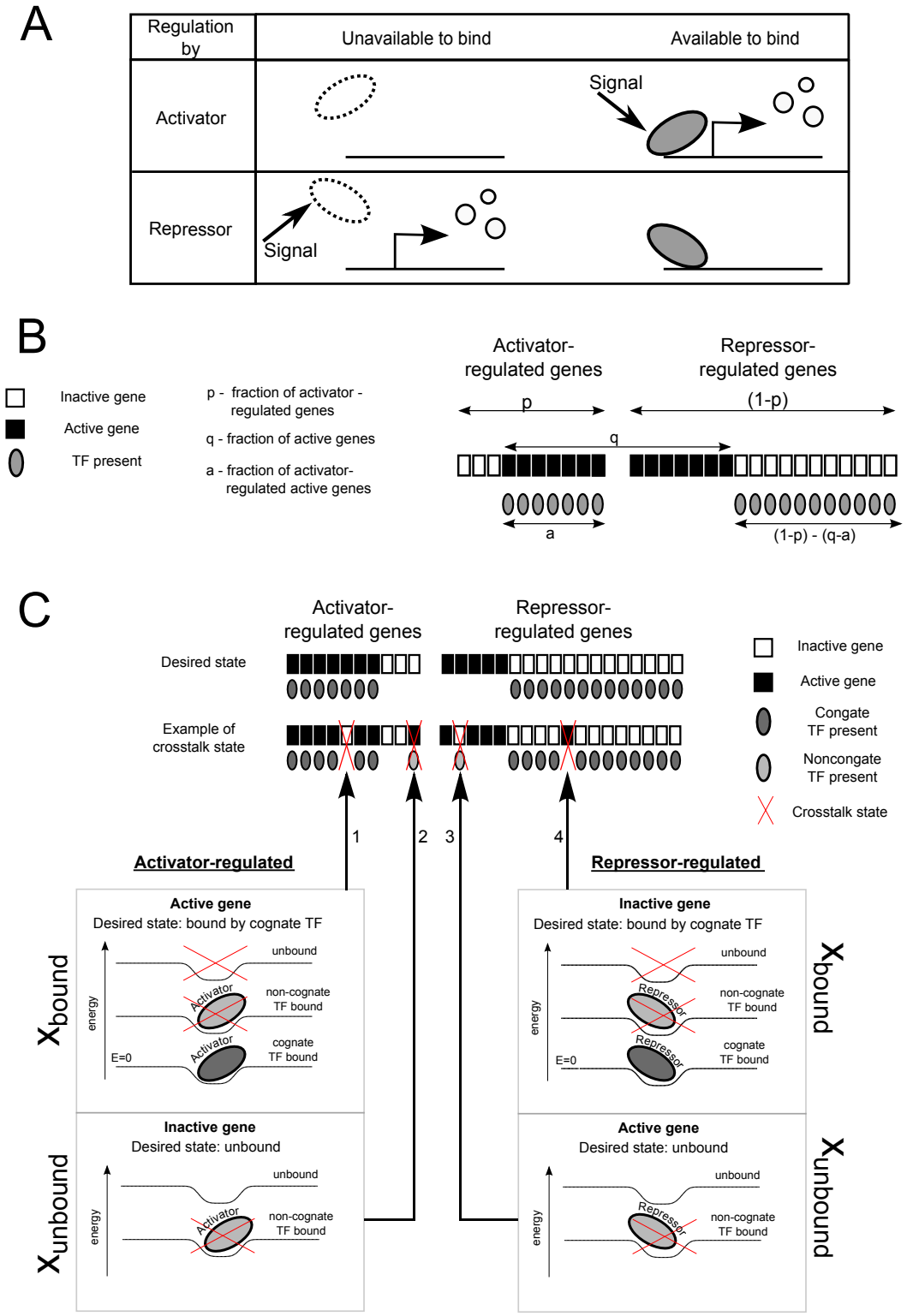

Figure 1: (Continued on the following page.)

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: **Gene regulation can employ different combinations of activators and repressors to implement the same gene expression pattern.** **(A)** A signal can cause gene activation by either positive (first row) or negative (second row) control. **(B)** We consider a total of M genes in a cell, of which a fraction $0 \leqslant p \leqslant 1$ is regulated by activators, and the remaining $1-p$ proportion is regulated by repressors. Assume that only a fraction $q < 1$ of these genes should be active under certain conditions (black squares), while the remaining genes should be inactive (white squares). In general, $a \leqslant q, p$ of this q proportion is activator-regulated and $q-a$ is repressor-regulated. Here, we illustrate all four cases of active/inactive genes regulated by activator/repressor and define all the variables. Gray ellipses represent TFs (of either type) required to maintain the regulatory state of the genes. **(C)** Different genes are regulated by different TF species, where TF specificity is determined by short regulatory DNA sequences (binding sites) adjacent to the gene. Each such binding site can be in different energy levels depending on its TF occupancy. It is in the lowest $E = 0$ (most favorable) level when bound by its cognate TF; it can be in a variety of higher energy levels if a non-cognate TF binds or if the site remains unoccupied (lower panel). In the upper panel, we illustrate the crosstalk-free 'desired state' (first row), where each TF binds its cognate target. Below (second row), we illustrate four different possibilities in which binding of a TF to non-cognate binding sites or failure to bind can lead to crosstalk: An activator-regulated gene should ideally be regulated by its cognate activator (right-inclined ellipse), in order to become active. If this cognate TF fails to bind when the gene should be active (1), or if another TF binds when the gene should remain inactive (2), we consider this as crosstalk. For a repressor-regulated gene, crosstalk states occur when a non-cognate repressor binds when the gene should be active (3), or if the cognate repressor fails to bind when the gene should be inactive (4). We present cognate TFs by dark gray and non-cognate ones by light gray. Activators are represented by right-inclining and repressors by left-inclining ellipses. Crosstalk states are red-crossed.

$$S_i \equiv \langle e^{-\epsilon d_{ij}} \rangle_{P(d)} = \frac{1}{C} \sum_{j \neq i} C_j e^{-\epsilon d_{ij}} = \frac{1}{T} \sum_{j \neq i} e^{-\epsilon d_{ij}}. \tag{1}$$

As only a subset of the genes is regulated, the summation of only the corresponding subset of TFs available to bind is taken. S_i is defined as the average of the Boltzmann factors, $e^{-\epsilon d_{ij}}$, taken over the distribution of mismatch values $P(d)$ between binding sites i and j, $\forall j$. In the last equality in Eq. (1), we assume that all available TFs are found in equal concentrations $C_j = C/T, \forall j$, where C is the total TF concentration and T is the number of distinct TF species available. We found that allowing different concentrations for activators and repressors does not reduce crosstalk below this equal concentration scheme (SI). We also assume full symmetry between binding sites i, such that $S_i = S \ \forall i$. A theoretical analysis of a more general case with unequal S_i values can be found in SI. The value of S can be either estimated using binding site data (see example in Box 1) or analytically calculated under

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

different assumptions on the pairwise mismatch distribution $P(d)$. In the following, we use rescaled variables: $s = S \cdot M$ for rescaled similarity between binding sites, the fraction of available TFs ($t = T/M$) and the rescaled total TF concentration ($c = C/M$).

We distinguish crosstalk states of genes whose desired state of activity requires unoccupied binding sites (x_{unbound}), and those requiring occupation by a cognate regulator (x_{bound}). x_{unbound} crosstalk includes the cases of an activator-regulated gene that should remain inactive as well as that of a repressor-regulated gene that should be active, both requiring an unoccupied binding site. For these genes, the cognate TF is not available to bind and any binding event by another (non-cognate) regulator is considered crosstalk. x_{bound} crosstalk includes both an activator-regulated gene that should be active and a repressor-regulated one that should be inactive. For these, crosstalk states occur either if the binding site remains unbound or if it is occupied by a non-cognate regulator, in which case, the regulatory state is not guaranteed. For illustration of all possible crosstalk states, see Fig 1C. Using equilibrium statistical mechanics, these crosstalk probabilities for a single gene i are [15, 16, 13]:

$$x_{\text{bound}} = \frac{e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{i,j}}}{C_i + e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{i,j}}} = \frac{e^{-E_a} + cs}{c/t + e^{-E_a} + cs} \tag{2a}$$

$$x_{\text{unbound}} = \frac{\sum_{j\neq i} C_j e^{-\epsilon d_{i,j}}}{e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{i,j}}} = \frac{cs}{e^{-E_a} + cs}. \tag{2b}$$

E_a is the energy difference between cognate bound and unbound states. The expression $\sum_{j\neq i} C_j e^{-\epsilon d_{ij}}$ captures the sum of all interactions of binding site i with foreign regulators.

Global crosstalk depends on the use of regulators

We define the global crosstalk, X of a cell as the average fraction of genes found in any of the crosstalk states. For a given value of a, we average over different choices of a active genes out of the p activator-regulated and $q-a$ out of the $(1-p)$ repressor-regulated proportions. The weighted sum over these four types of contributions provides the average total crosstalk,

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

X, of the whole system:

$$\begin{aligned} X &= \overbrace{a \cdot x_{\text{bound}} + (p-a) \cdot x_{\text{unbound}}}^{\substack{\text{Contribution of} \\ \text{activator regulated genes}}} + \overbrace{(q-a) \cdot x_{\text{unbound}} + (1-p-q+a) \cdot x_{\text{bound}}}^{\text{Contribution of repressor regulated genes}} \\ &= t \cdot x_{\text{bound}} + (1-t) \cdot x_{\text{unbound}}. \end{aligned} \tag{3}$$

As Eq. (3) shows, X simply depends on the fraction of available TF species $t = 1-p-q+2a$, where $t = T/M$, regardless of their role as activators or repressors. Importantly, global crosstalk does not directly depend on the fraction of active genes q. This is a generalization of the result obtained in [13], where the special cases of $t = q$ (all TFs are activators) and $t = 1-q$ (all TFs are repressors) were studied. To obtain a lower bound on crosstalk values for given similarity, s, and fraction of available TFs, t, we substitute the expressions for x_{bound} and x_{unbound} (Eq. (2)) into Eq. (3). We then minimize X with respect to the total TF concentration, c, to obtain the expression for minimal crosstalk:

$$X^*(t,s) = t\left(-s(1-t) + 2\sqrt{s(1-t)}\right). \tag{4}$$

Hence, the lower bound on crosstalk X^* only depends on two macroscopic variables: s (similarity between binding sites) and t (fraction of available TFs). The higher the similarity s is, the larger is the resulting crosstalk X^*, where to first order, $X^* \sim \sqrt{s}$ (Fig 2A). The dependence on t is more complicated and non-monotonic: for low t values $t < t^*(s)$ (we show in the SI that $t^*(s) \geqslant 2/3$), X^* increases with t. Intuitively, the number of TF species present positively correlates with the number of crosstalk opportunities. Contrary to this intuition, for high TF usage beyond the threshold value t^*, we find the opposite trend, where X^* *decreases* with increasing TF usage, t. This non-monotonic dependence of X^* on t comes about because global crosstalk balances between binding sites that should be bound for which higher c reduces crosstalk and binding sites that should be unbound, for which an increase in c has an opposite effect. The relative weight of binding sites that should be bound vs. those that should be unbound shifts with t. High TF usage always comes at the cost of an exponential increase in the optimal TF concentration, c^* (see Eq. S4 for expression) needed to minimize crosstalk, where for high s values, c^* diverges to infinity

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$c^* \to \infty$ (see Fig 2B). We discuss below the biological relevance of this regime.

Mode of regulation affects global crosstalk because it affects TF usage

A gene activity pattern can be obtained by different combinations of positive and negative regulation, yielding seemingly identical gene functionality. One may then ask whether these various TF-gene associations differ in the resulting global crosstalk. Following Eq. (4), crosstalk only depends on the fraction of available TF species, t, regardless of the underlying association of a gene with either activator or repressor. It is thus sufficient to consider how different regulatory strategies affect TF usage, rather than analyzing the whole network architecture, thereby significantly simplifying the analysis. Using our model, we calculate the global crosstalk for any combination of the fraction of active genes, q, with any mixture of activators and repressors defined by p. That covers all possible gene-regulator associations with either activators or repressors. While each point represents a fixed fraction of active genes, this model can also be used to study a varying number of active genes, by taking a distribution of points over the q-axis (see SI for an example). Specifically, we focus on the two extreme gene-regulator associations, which we call the 'busy' and 'idle' strategies. The 'busy' strategy means that gene regulation is operative most of the time. It is implied by the "Savageau demand rule" [2], because the genes' default state of activity is not their commonly needed state. Under the opposite 'idle' strategy, the default state of each gene is its more commonly needed regulatory state. Hence, regulation is inoperative most of the time (see Fig 2C). Hybrids of these two extreme strategies are also possible.

To represent the 'busy' strategy, we associate as much as the q active proportion as possible with activators, and only if the total fraction of activators is smaller than the fraction of active genes ($p < q$), the remaining $q - p$ proportion will employ regulation by repressors. Thus the fraction of activator-regulated active genes equals $a = \min(p, q)$. Conversely, under the 'idle' strategy, we associate as much as the q active proportion as possible with repressors. Only if the fraction of repressors is smaller than the proportion of active genes ($1 - p < q$), the remaining active genes will pursue positive regulation, hence $a = q - \min((1 - p), q)$. The corresponding fractions of TFs in use (including both activators and repressors) in these two extremes are then:

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

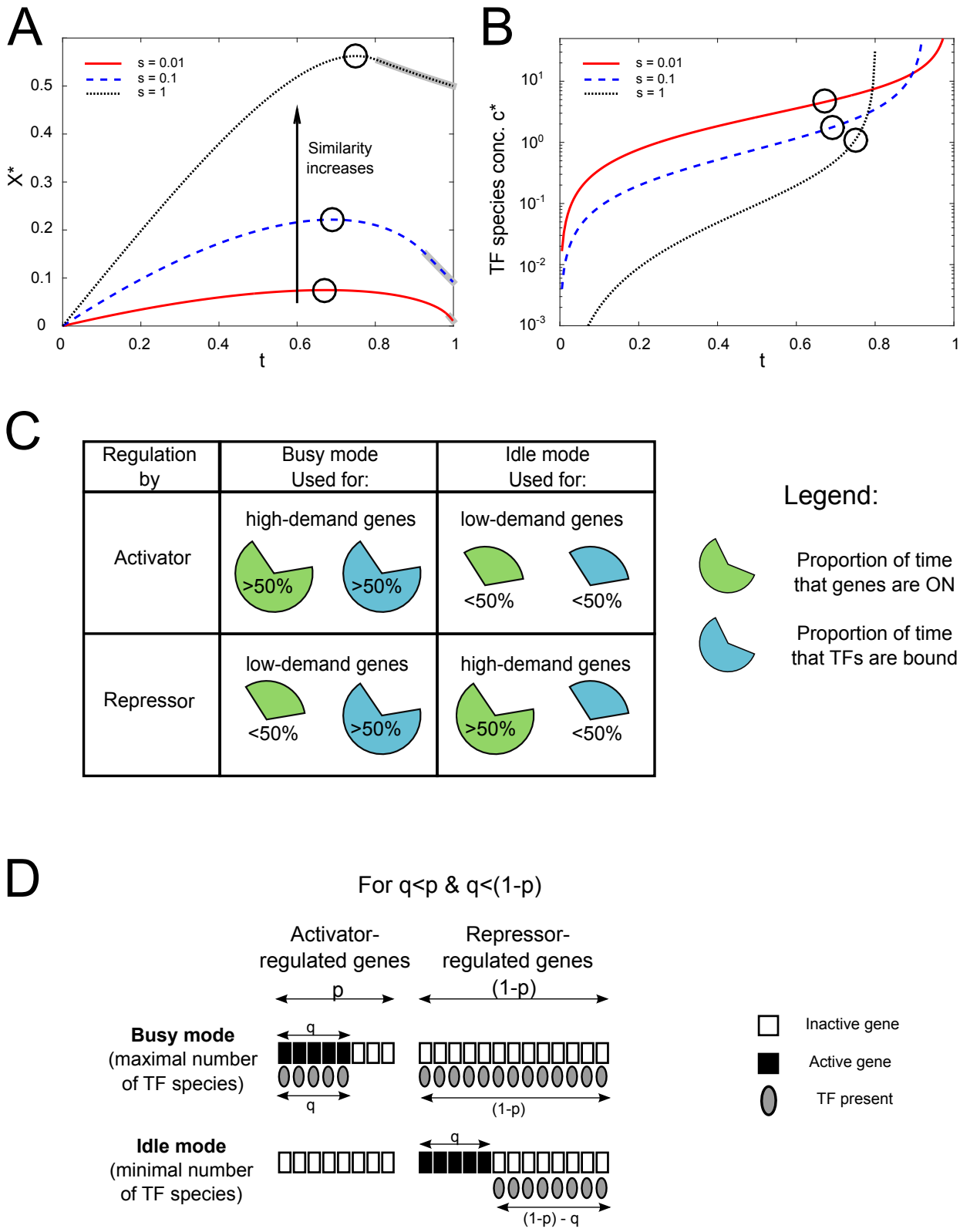

Figure 2: (Continued on the following page.)

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2: **Crosstalk depends on the fraction of TFs in use, which varies between regulatory strategies.** **(A)** Minimal crosstalk, X^* vs. t the fraction of TFs used for different values of similarity, s, between distinct binding sites. In most of the parameter regime (for $t < t^*$, $t^* \geqslant 2/3$), minimal crosstalk, X^*, increases with t. Black circles denote the maxima of the curves. Crosstalk monotonically increases with similarity between binding sites. The anomalous regime where TF concentration needed to minimize crosstalk mathematically diverges to infinity, is grey-shaded around the curves. **(B)** The optimal TF concentration, c^*, needed to minimize crosstalk increases sharply with t. c^* diverges to infinity at the boundary with the anomalous regime, which for high similarity s, occurs already at lower TF usage t. Circles represent the maximal X^* values for each curve (as in (A)). **(C)** Different genes are expressed to different extents, where here, we grossly classify them as either high- (more than half of the time) or low-demand (less than half). If a high-demand gene is regulated by an activator or if a low-demand gene is regulated by a repressor, demand for the regulator will be high ('busy mode'). Conversely, if the same high-demand gene is regulated by a repressor and the low-demand gene is regulated by an activator, the regulator is only required a small fraction of the time ('idle mode'). **(D)** Each of the q active genes and $1-q$ inactive genes can be assigned either positive or negative regulation. We illustrate the two extremes maximizing (minimizing) TF usage: in the 'busy' ('idle') mode, as many active genes as possible are assigned positive (negative) regulation and as many inactive genes as possible are assigned negative (positive) one. The scheme is shown for an example with the given relationship between the proportion of active genes q, the proportion of activator-regulated p and the proportion of repressor-regulated $(1-p)$; that is $q \leqslant p, (1-p)$. Other combinations are shown in SI Fig S5.

$$t_{\text{busy}} = 1 - |p - q| \tag{5a}$$

$$t_{\text{idle}} = |1 - p - q|. \tag{5b}$$

In Fig 2D, we illustrate regulation following these two extreme strategies. We show that the TF assignments defined in Eq. (5) are the two extremes in TF usage: namely, for any general regulatory scheme, the fraction of TFs needed to regulate a given fraction of genes q is $t_{\text{idle}} \leqslant t \leqslant t_{\text{busy}}$ (see SI for formal proof).

In Fig 3A, we illustrate the difference in the fraction of available TFs between the two extreme strategies $\Delta t = t_{\text{busy}} - t_{\text{idle}} = 1 - |p - q| - |1 - p - q| > 0$, demonstrating that the 'busy' strategy always requires more regulators than 'idle' strategy (see SI).

Using Eq. (4), we obtain exact expressions for X^* under these extreme strategies (see SI). In Fig 3B, we show $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$, the difference in minimal crosstalk X^* between the two extreme strategies for all combinations of the fraction of active genes $0 \leqslant q \leqslant 1$

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and fraction of TFs that are activators $0 \leqslant p \leqslant 1$. We find that the 'idle' mode yields less crosstalk in a large part of this parameter space. The 'busy' mode still involves less crosstalk for parameter combinations centered around the diagonal $p = q$, whereas the 'idle' strategy always performs best on the anti-diagonal $1 - p = q$. This is because, on the diagonal, the fraction of activators, p, equals exactly the fraction of genes that should be active (q) resulting in full usage of all existing TFs, $t = 1$. On the anti-diagonal $1 - p = q$, the fraction of genes that should be active, q, equals exactly the fraction of repressors $1 - p$. Thus the default state of all genes is the desired regulatory state requiring no TF usage at all, $t = 0$, which makes the 'idle' strategy most advantageous.

The region in which 'busy' mode yields the lowest crosstalk comes at the cost of using a larger fraction of existing TF species, as depicted in Fig 3C. The 'idle' mode, in contrast, requires a much smaller fraction of TF species to be in use. Furthermore, the two strategies differ not only in the fraction of TFs needed but also in their concentrations. To achieve the lower bound, the 'busy' strategy always requires a total higher TF concentration, c^* (Fig 3D).

The explanation for the alternating crosstalk advantage between the two extreme strategies lies in the non-monotonic dependence of crosstalk on TF usage, t (Fig 2A). For $t(p, q) < t^*(s)$, crosstalk *increases* and for $t(p, q) > t^*(s)$ it *decreases* with t. Thus, for (p, q) combinations for which $t_{\text{idle}} < t_{\text{busy}} < t^*$, 'idle' strategy will yield lower crosstalk, whereas if $t^* < t_{\text{idle}} < t_{\text{busy}}$ 'busy' will be more advantageous in that sense (see SI for more details). While 'idle' and 'busy' represent the two extremes, a continuum of regulatory strategies interpolating between these two extremes can be defined. We show, however, that minimal crosstalk is always obtained by one of the two extremes, due to the concavity of $X^*(t)$ (see SI).

We previously found that for some parameter combinations of similarity, s, and fraction of active genes, q, the mathematical result of the lower bound on crosstalk X^* (Eq. (4)) has no biological relevance [13]. Specifically, for similarity between binding sites which is too high $s > \frac{1}{1-t}$, regulation is ineffective and the lower bound on crosstalk X^* is obtained with no regulation at all. For high TF usage $t > t_{\text{max}}$ (see SI of [13]), the concentration needed to obtain minimal crosstalk formally diverges to infinity $C^* \to \infty$, which is again a biologically

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

irrelevant regime. These biologically implausible regimes put an upper bound to the total number of genes that an organism can effectively regulate [25, 13]. The results shown in Fig 3 only refer to crosstalk values obtained in the 'regulation regime' where C^* is finite and positive $0 < C^* < \infty$. Specifically, we find that when similarity, s, increases, parts of the parameter space shown in Fig 3A indeed move into the anomalous regimes. In particular, the high TF usage region around the diagonal $p = q$, where the 'busy' strategy outperforms in crosstalk reduction, vanishes due to this anomaly (see Fig 3E where anomalous regions are blackened). For high similarity values $s > 5$, the 'idle' strategy yields lower crosstalk in the entire biologically relevant parameter space - see SI and Fig S7.

The distribution of crosstalk in a stochastic gene activity model

So far, we considered a deterministic model in which the numbers of active genes and available TF species were fixed, resulting in a single crosstalk value per (p, q) configuration. In reality, these numbers can temporally fluctuate, for example, because of the bursty nature of gene expression [26, 27]. In the deterministic model, we also assumed uniform gene usage, such that all genes are equally likely to be active. In reality, however, some genes are active more frequently than others.

To account for this, we study crosstalk in a probabilistic gene activity model. We assume independence between activities of different genes, where each gene i has demand (probability to be active) D_i. We then numerically calculate crosstalk for a set of genes. This approach enables us to incorporate a varying number of active genes and a non-uniform gene demand and compare our results to the deterministic model studied above. To comply with its demand D_i, each gene i is regulated with probability γ_i, $i = 1...M$, where $\gamma_i = D_i$ if regulation is positive and $\gamma_i = 1 - D_i$ if it is negative. In Fig 3F, we demonstrate this calculation, for two values of t, representative of the two extreme strategies. We find excellent agreement between this numerical calculation and the analytical model studied above. The distribution of X^* is typically narrow, such that for practical purposes, the distribution mean, calculated using the deterministic activation model, serves as an excellent estimator of crosstalk values. For more details on this calculation and for approximation of the distribution width, see SI.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

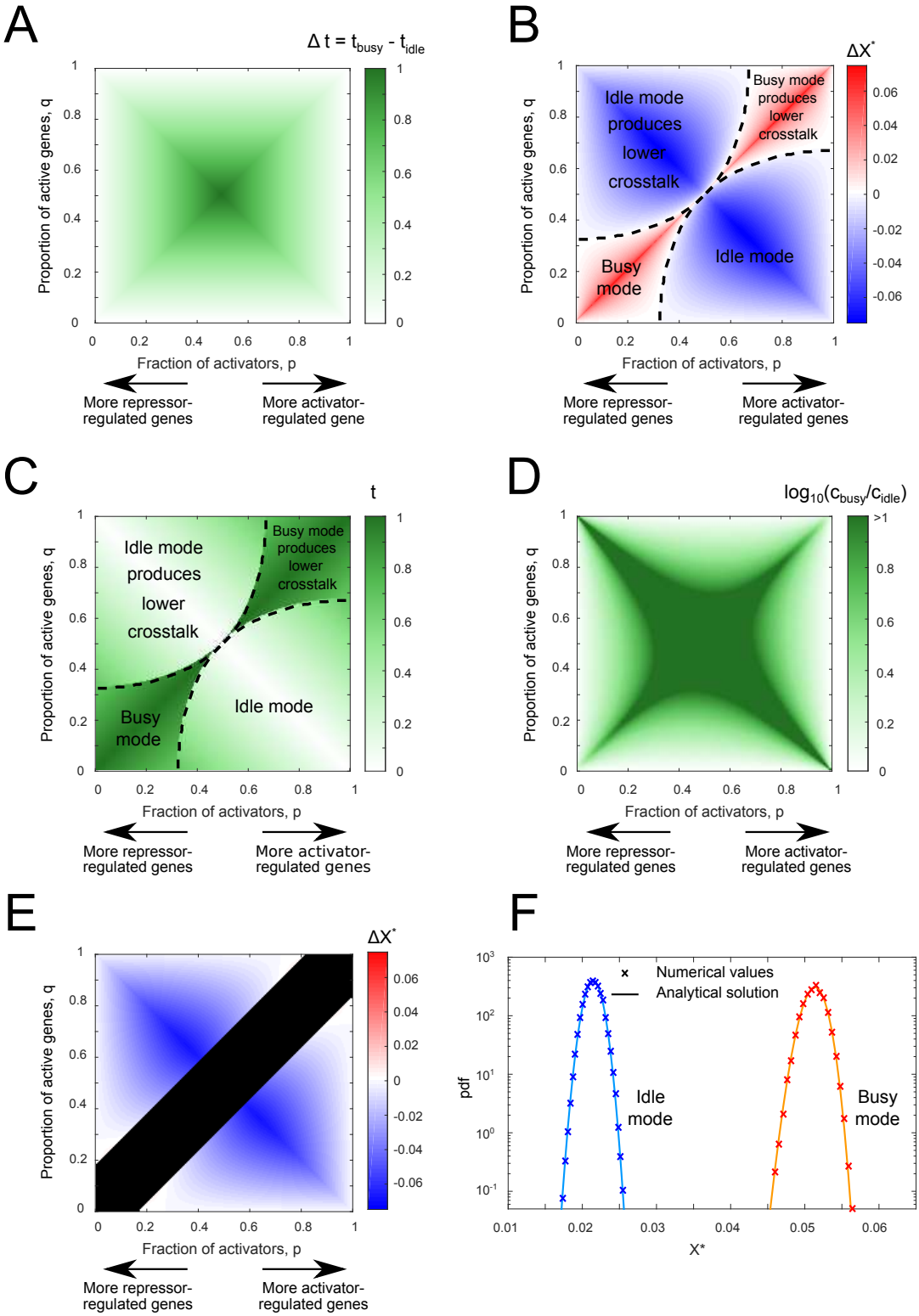

Figure 3: (Continued on the following page.)

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3: **Idle strategy yields lower crosstalk than 'busy' in a large part of the parameter regime.** **(A)** The 'busy' strategy always requires a larger TF usage compared to the 'idle' one. Here we illustrate Δt the difference in the fraction of TFs in use between the two strategies for different values of p and q shown in color scale. **(B)** The difference in total crosstalk ($\Delta X = X^*_{\text{idle}} - X^*_{\text{busy}}$) between 'idle' and 'busy' strategies, shown in color scale, as a function of the fraction of activator-regulated genes p and relative number of active genes q. In a large part of the parameter regime (colored blue) lower crosstalk is achieved by following the 'idle' strategy. The 'busy' strategy is most beneficial on the diagonal $p = q$ (red region), but this requires a use of all TFs and comes at the cost of enormously high TF concentration. The 'idle' strategy is most beneficial around the anti-diagonal $q \approx 1 - p$, where regulation can proceed with no TFs at all and crosstalk is close to zero. **(C)** Fraction of TFs in use when the strategy providing minimal crosstalk ('idle' or 'busy' as in (b)) is used, as a function of p and q shown in color scale. Black dashed lines mark the borders between the regions where 'busy' or 'idle' strategies provide lower crosstalk. While 'idle' strategy mostly requires a minority (<50%) of the TFs, the 'busy' one always necessitates a majority (>50%) of TFs to be in use. $s = 10^{-2}$ was used in (B)-(C). **(D)** Ratio between TF concentrations providing minimal crosstalk in either strategy $C^*_{\text{busy}}/C^*_{\text{idle}}$. 'Busy' strategy always requires higher TF concentrations. **(E)** For higher similarity s between binding sites parts of the parameter space falls into the anomalous regime where the optimal TF concentration diverges to infinity. We plot here the difference in optimal crosstalk $\Delta X = X^*_{\text{idle}} - X^*_{\text{busy}}$ between strategies for $s = 1$. Black areas denote the anomalous regime. Importantly, the region where the 'busy' strategy was beneficial for low s (see (B)) falls into this anomalous regime. **(F)** Numerical calculation of the distribution of crosstalk values, for different choices of TF usage, is in excellent agreement with simulation results. The distributions obtained are narrow, suggesting that their mean value is representative. Crosstalk values only depend on TF usage, regardless of the exact underlying model. Parameter values: total number of genes $M = 3000$, proportion of activator regulated genes $p = \frac{1}{3}$, regulation probability $\gamma_i = \gamma = 0.12$ for 'idle' mode and $\gamma_i = \gamma = 0.92$ for 'busy' mode with $2 \cdot 10^6$ realizations.

Discussion

The existence of multiple system designs that seemingly realize the same function, is ubiquitous and often puzzling [28, 29, 30, 31]. Naturally, it raises two types of questions [2]. Are there secondary functional differences between different designs? Do such designs differ in their evolutionary accessibility (and sustainability) [32, 33] or is the existence of one or the other just a coincidence [11]? Here, we concentrated on a particular functional aspect of gene regulatory networks: crosstalk in transcriptional regulation. Using a mathematical model for global crosstalk, we study crosstalk limits under different gene regulatory network designs that implement the same gene activity pattern in response to a signal. We found a lower bound on crosstalk $X^* = X^*(t, s)$ [13], which is fully determined by two

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

macroscopic "thermodynamic-like" variables, regardless of other microscopic details of the network. These are the fractions of available TF species, t, and the average similarity between distinct binding site sequences, s. This emergent simplification enables us to analyze crosstalk for a general gene regulatory network, regardless of its exact design.

We found that minimal crosstalk has a non-monotonic dependence on the fraction of available TF species, t: in a large part of the parameter regime, the minimal achievable *global* crosstalk increases with the fraction of TFs in use. Only at the very high end of t values and if binding sites are well-distinguishable from each other ($s \ll 1$), this trend is reversed. Hence, in a large part of the parameter regime, crosstalk is simply minimized by minimization of TF usage. The latter is minimized if genes are, by default, in the regulatory state required most often, namely, genes whose product is often needed ('high-demand' genes) are by default active and employ negative regulation to become inactive, while genes whose product is only sporadically needed ('low-demand') are by default inactive and pursue positive control to be active. We called such a gene regulation scheme the 'idle' strategy. This is opposed to the 'busy' strategy, which implements the 'Savageau demand rule' [1], which consequently *maximizes* TF usage. While the 'busy' regulatory strategy still minimizes crosstalk in a small part of the parameter space, it always carries the additional cost of a much higher total TF concentration required. The advantage of one regulatory strategy over the other also depends on the similarity between binding sites, s. We found that the portion of the parameter space in which the 'busy' strategy provides lower crosstalk shrinks when the similarity, s, increases, until it vanishes completely for high similarity $s > 5$ (see SI). As the fractions of co-active genes q and activators p constrain the possible minimal and maximal TF usage values, t, the value of minimal crosstalk and the crosstalk advantage of either regulatory strategy also vary with q and p. In general, the 'busy' strategy performs best along the diagonal $p = q$, where the fraction of active genes exactly equals the fraction of activators, while the 'idle' strategy performs best on the anti-diagonal $p = 1 - q$, where the fraction of active genes equals the fraction of repressors.

Where are real organisms located in this parameter space? Reports of the number of co-expressed genes greatly vary between organisms and strongly depend on growth conditions. For example: $\sim$ 10,000 different genes were reported to be co-expressed in a mouse cell

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(<50% of total) [34, 35], 10,000-12,000 genes were estimated to be co-expressed in human HeLa cells [36], in *E. coli* during exponential growth, 76%-82% (3300-3500 out of 4290 genes) were expressed [37, 38] and in *S. cerevisiae*, 75%-80% of the genes were expressed [39, 40].

Values of similarity between binding sites, s, vary not only between organisms, but also between modules and distinct genes within the same organism (see Box 1). We estimated s and the resultant minimal crosstalk values for *S. cerevisiae* using binding site data of 23 TFs. We found a broad distribution of s values spanning 5 orders of magnitude, which peaked around $10^{-4} - 10^{-3}$. For such low s values, there is a regime in parameter space in which 'busy' yields the lowest crosstalk. We found that if 75% of the genes are simultaneously expressed, the 'busy' mode should provide lower crosstalk if more than 60% of the TFs are activators (where otherwise lowest crosstalk would be provided by 'idle'). Our estimates for minimal crosstalk X^* for this sub-network of *S. cerevisiae* were in the range 0.02-0.06 (see Box 1), assuming 75% of the genes are co-regulated. However, these estimates may be biased, as they are based only on a small number of TFs, each regulating a large number of targets.

Lastly, we studied a stochastic gene activation model where the number of co-active genes was a random variable and X^* took a distribution of values. We found that the resultant TF availability and minimal crosstalk distributions are typically very narrow and centered around their mean value. Hence, for practical purposes the deterministic activation model which only provides the mean value, is an excellent approximation.

The significance of our analysis is that it provides a lower bound. Some of our simplifying assumptions are lenient and crosstalk could be higher than predicted by our model. For example, we assumed that total TF concentration is accurately tuned. In reality, TFs are not necessarily expressed and degraded at a precise time [41], potentially leading to higher crosstalk levels than predicted by our model. Other simplifications that might affect crosstalk estimates include the averaging over gene sets of same-size q as representatives of different environmental conditions, whereas, in reality, the number of expressed genes could vary between environments (e.g., growth media [37]). We also averaged over all possible choices of q active genes, but only some of these activity combinations occur naturally. Furthermore, Hershberg and co-workers found that during evolutionary processes, repressor-

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regulated genes are typically lost before their repressor is lost, but activators are commonly lost before their downstream genes are lost [42]. An interesting consequence of this finding is an imbalance between genes and regulators, where orphan repressors with no genes and orphan genes with no activators, transiently exist, in contrast to our idealized model where every gene has a regulator, and vice versa. These excess genes and repressors can still be functional and contribute to global crosstalk.

Here, we focused on the crosstalk implications of activator vs. repressor regulation. Additional considerations can be involved in network evolution and potentially tip the balance between different regulatory architectures observed in nature. For example, a previous theoretical study [28] found that evolutionary divergence of repressors requires higher selection pressures compared to divergence of activators. This consideration could bias the TF composition in favor of more activators than repressors. Crosstalk was suggested there to play an important evolutionary role, facilitating the evolution of novel network structures, as the functionality of intermediates relies on TF low specificity [28]. A network's robustness to mutations vs. its potential to evolve new forms, have been highlighted as evolutionary determinants [33]. This calls for revisiting evolution of gene regulation models, to account for global network effects, and generalize the previously studied single-gene models.

In sum, our study demonstrated the pervasive impact of a new type of protein production cost which is typically overlooked: that of regulatory interference caused by excess proteins in the dense cellular medium. This comes on top of the energetic burden of unnecessary protein production, which was found to delay growth [8, 9, 10, 43]. The regulatory cost is predicted to increase super-linearly with the number of molecular species and interactions involved. Hence, it only becomes apparent, once the network is considered as a whole. Here, we propose that their cumulative effect could bias the choice between regulatory topologies. Indications of selection to reduce such regulatory interference were indeed found in genomic data [44, 45], and similar costs were proposed for protein-protein interactions [46, 47] and DNA self-assembly [48]. We leave it for future work to construct a generalized model incorporating evolutionary as well as biophysical considerations of the cost of complexity in biological networks, in general, and regulatory architectures, in particular.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Box 1

In our theoretical model, we have several simplifying assumptions to allow for an analytical solution. In particular, we assumed uniform properties for all binding sites, assigned equal energetic contributions to all nucleotides in the sequence and assumed one-to-one association of TF to downstream gene. These assumptions do not generally apply, hence, application of our model to real gene regulatory networks will require a data-based numerical calculation of crosstalk. Here, we demonstrate the first steps in developing the necessary computational techniques and discuss the concomitant technical challenges. Specifically, we relax some of our simplifying assumptions in a numerical calculation of crosstalk using genomic data of binding site sequences. An exhaustive measurement of all TF concentrations is technically possible, but such data does not yet exist, to the best of our knowledge. Once it is available, the calculation of actual crosstalk values will, in principle, be possible, yet will require an enormous number of computations to account for all binding possibilities of each TF to each possible binding site [49]. Such calculations have, so far, only been applied to small-scale networks of only few genes [14, 50]. Hence, we still employ the assumptions of TF concentrations that are either equal or proportional to the number of genes regulated by each TF. Due to lack of data on the effects of combinations of mutations, we still assume additivity in the energetic contributions of different positions in the sequence. This assumptions is considered reasonable for up to 3-4 bp substitutions [21]. We use binding site sequence data of various transcription factors, provided as position count matrices (PCMs). PCMs are $4 \times L$ matrices that provide the total number of counts for each nucleotide at each of the L binding site positions, taken over multiple binding sites of the particular transcription factor. They allow us to compute the mismatch energy penalties for every position and nucleotide in a given binding site sequence. We numerically calculate crosstalk using PCM data for 23 *S. cerevisiae* transcription factors collected from scerTF database [51, 52].

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similarity values vary between genes even within the same organism

Our numerical calculation of total crosstalk follows the thermodynamic model of gene expression. We randomly chose which TFs are available and then computed the probabilities for all possible TF-BS binding states. We calculated the between-binding site similarity s using the TF binding sequence data (see Methods). Only binding states associated with transcription factors we randomly chose as available, were taken into account in this calculation - rather than including all TF contributions, as we have done in the analytical solution. In Fig 4A, we show the distribution of similarity values of genes associated with 23 *S. cerevisiae* transcription factors. We find a broad distribution of s values spanning over 5 orders of magnitude, peaking at $10^{-4} - 10^{-3}$. This finding is in contrast to the fully symmetric picture obtained with equal s values for all genes in our analytical solution.

Fig 4B shows the comparison between numerical (solid black line) and analytical (dotted black line) total crosstalk values. In this numerical calculation, we assumed that all genes regulated by each TF are lumped together into a single "effective gene". In the analytical calculation, we assumed a single s value common to all binding sites. The numerical prediction exceeds the analytical one, which could be attributed to the broad range of similarity values, where contributions from few genes with high s could dominate crosstalk.

Incorporation of a complex TF-gene interaction network

Thus far, we assumed that each TF regulates only a single unique gene. Yet, in real gene regulatory networks, the same TF species often regulates multiple genes and some genes are regulated by a combination of different TFs. To account for this, we expanded our dataset to include all the 2126 genes [53] regulated by the 23 *S. cerevisiae* transcription factors and considered all TF-gene interactions in this set.

Here, we assumed that TF concentrations are proportional to the number of genes they regulate. As similarity is a function of TF concentrations, this modification also

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

affects the single similarity value used for analytical prediction. Fig 4B shows both the numerically calculated crosstalk and the analytically predicted one (with modified s) for this more complex interaction network (solid and dashed gray lines).

More details about the numerical calculation are found in SI.

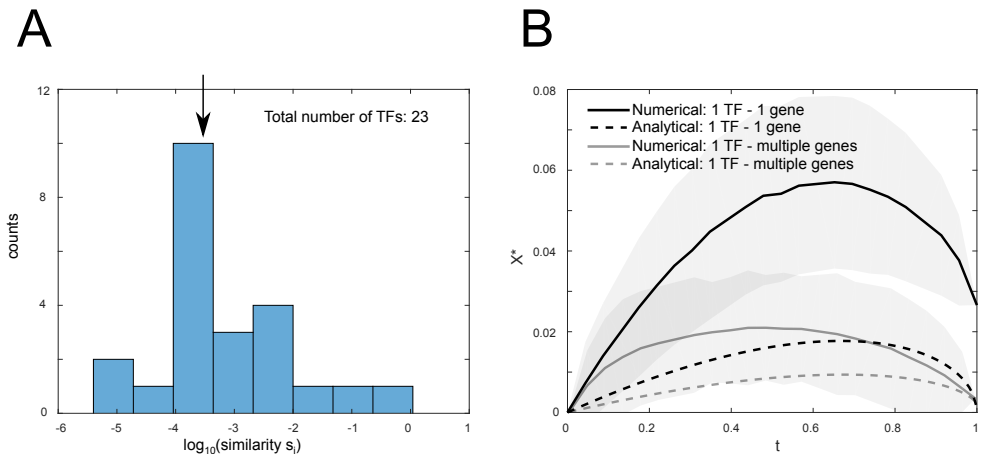

Figure 4: **Data-based crosstalk estimates.** **(A)** Per-gene similarity values for *S. cerevisiae* exhibit a broad distribution spanning five orders of magnitude with median value of $3 \cdot 10^{-4}$ (arrow). This calculation is based on binding site data of 23 *S. cerevisiae* TFs [51, 52, 53] (see Methods). We assume a single "effective gene" that lumps together all downstream genes regulated by each TF. We also assume that all TF species are simultaneously available to bind (and cause crosstalk). **(B)** Numerical prediction of global crosstalk for *S. cerevisiae* (solid lines) compared to analytical ones based on a single similarity value common to all genes (dashed lines) taken as the median value of the single gene similarity measures s_i. The black curves show calculation assuming that each of the TFs regulates only one gene (1 TF - 1 gene) while gray curves represent estimation of crosstalk for the network of all 2126 *S. cerevisiae* genes regulated by these 23 TFs (1 TF - multiple genes). Usage of all downstream genes yields lower crosstalk than the single-gene approach, since the majority of genes are regulated by TFs associated with low similarity values. The numerical calculation however yields higher crosstalk than the analytical one, since it is dominated by the few high similarity TFs, which are further away from the median. The numerical curves represent the mean over 10^4 realizations for each t where the subset of available TFs was randomly drawn at each. The surrounding gray shadings show ± 1 standard deviation around the mean.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Distribution of t is approximated by a Gaussian distribution. Given that a gene i has its cognate TF present with a probability γ_i ($i \in (1, M)$, where M is the total number of genes), the distribution of available transcription factor species in the system follows Poisson-binomial distribution. This is the probability distribution of a sum of independent Bernoulli trials that are not necessarily identically distributed, each with probability γ_i. Its mean and variance are,

$$\langle t \rangle = \frac{1}{M} \sum_{i=1}^{M} \gamma_i = \langle \gamma_i \rangle, \tag{6}$$

$$\mathrm{var}(t) = \frac{1}{M} \sum_{i=1}^{M} \gamma_i (1 - \gamma_i) = \langle \gamma_i (1 - \gamma_i) \rangle. \tag{7}$$

As this distribution is difficult to compute for large values of M, we can follow the central limit theorem and approximate it by a Gaussian distribution with the same mean and variance.

Exact solution of probability distribution of X^*. For a function $X^*(t)$, where t is a random variable with probability distribution $f_t(t)$, the probability distribution of X^*, $f_{X*}(X^*)$ is:

$$f_{X*}(X^*) = \sum_i f_t(g_i^{-1}(X^*)) \left| \frac{\mathrm{d}g_i^{-1}(X^*)}{\mathrm{d}X^*} \right|, \tag{8}$$

where $g_i^{-1}(X^*) = t_i$ represents the inverse function of the $i-$th branch. In our case with two branches :

$$f_{X*}(X^*) = f_t(g_1^{-1}(X^*)) \left| \frac{\mathrm{d}g_1^{-1}(X^*)}{\mathrm{d}X^*} \right| + f_t(g_2^{-1}(X^*)) \left| \frac{\mathrm{d}g_2^{-1}(X^*)}{\mathrm{d}X^*} \right|. \tag{9}$$

The solutions for $g_i^{-1}(X^*)$ and its derivative exist for crosstalk $X^*(t)$ and can be analytically computed. Therefore, the solution for the distribution of minimal crosstalk $f_{X*}(X^*)$ is also analytically known.

For regime I, the lower limit on crosstalk is $X^*(t) = t$. Its inverse is $g^{-1}(X^*) = t(X^*) = X^*$,

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

while the derivate $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = 1$. Similarly, in regime II, the lower limit of crosstalk equals $X^*(t) = 1-t/(1+\alpha t)$, the inverse function $g^{-1}(X^*) = t(X^*) = (1-X^*)/(1-\alpha+\alpha X^*)$, and its derivate $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = -(1-\alpha-\alpha X^*)^{-2}$. The analytical solution for regime III was computed using Mathematica and the solution can be found in SI.

Using these values, one can compute $f_{X^*}(X^*)$ for X^* coming from any of the three regimes.

Stochastic semi-analytical solution of crosstalk for random number of TFs present. For each gene i, we draw by random, with probability γ_i, whether its cognate TF is available. We then obtain the proportion t of available TFs. As this process is stochastic, the proportion t differs between different realizations. Next, we compute the lower limit on crosstalk $X^*(t)$ for this t value using the analytical solution. The crosstalk is computed in regime I, II or III, as needed. Using multiple realizations ($=10^6$) of t, we numerically obtain the distribution of crosstalk values for values of $t \in (0, 1)$.

Obtaining the energy matrices from position count matrices (PCMs). Position count matrices (PCMs) document the binding site sequences of TFs. Each element c_{ij} designates the number of known TF binding site sequences with nucleotide i in position j. We obtain the PCMs from scerTF database for *S. cerevisiae*. Given these, we calculated the energy matrices which are needed to compute the similarity measure, in the following way. For a position j and nucleotide $i \in \{A, C, G, T\}$, we compute the energy mismatch value as $\epsilon_{ij} = \log\left(\frac{c_{mj}}{c_{ij}}\right)$, where c_{ij} is the count of nucleotide i in position j and $c_{mj} = \max_i c_{ij}$ is the nucleotide at the position j with the maximal count. In case of zero counts, $c_{ij} = 0$, where the energy ϵ_{ij} diverges, we add a constant pseudocount $\delta = 0.1$ to matrix entries.

Numerical computation of similarity measure using PCMs. To compute the similarity measure between binding site k and a transcription factor l, we first substitute the sequence of BS k by the *consensus* sequence of its cognate TF k. The consensus sequence is obtained by taking the most common nucleotide in each position j. As the given binding site and TF consensus sequence are not necessarily of the same length, we distinguish between the following cases:

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- If the TF consensus sequence l is shorter than the binding site sequence k, we compute the energies for all possible overlaps of the shorter sequence with respect to the longer one. We take the minimal value to be the binding energy.

- If the TF consensus sequence l is longer than the binding site sequence k, the TF energy matrix is again slid over the binding site and energies again are calculated for every relative positioning of the two sequences. The only difference from the previous case is that energetic contributions from positions where the TF binds outside the binding site are taken into account by averaging energies over all four nucleotides. In other words, total binding energy $E = E_1 + E_2$ is the sum of contributions from nucleotides inside (E_1) and outside (E_2) the binding site. The energy contribution of positions j outside of BS equals $E_2 = \sum_j E_{2j}$, with $E_{2j} = \sum_{i=1}^{4} \epsilon_{ij}/4$ being the average binding energy at position j (ϵ_{ij} represent elements in the energy matrix). Here too, we compute the binding energy for all possible overlaps between BS and TF. Then, the lowest energy is taken as the binding energy E_{kl}.

This provides the matrix of binding energies E^{kl} between every binding site k and every TF l. Importantly, this binding energy is asymmetric, namely $E^{kl} \neq E^{lk}$. Hence, the similarity measure between binding site k and all other binding sites is computed as:

$$S_k = \sum_{l=1,\ l\neq k}^{M} C_l e^{-E^{kl}}, \tag{10}$$

with C_l being the concentration of TF species l. In other words, the similarity S_k is the average Boltzmann weight, taken over all non-cognate TF binding to binding site k.

Numerical computation of crosstalk given PCMs. For the numerical computation of crosstalk, we use the matrix of binding energies E^{kl} between binding site k and TF l. The procedure is as follows:

1. randomly choose a subset t (e.g., $t = 0.2$ represents a fifth of all TFs) of all TFs that should be regulated by their cognate TF. With each realization, a different subset is chosen.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. for gene k, obtain the similarity measure $S_k = \sum_{l \neq k} C_l e^{-E^{kl}}$, with C_l being the concentration of TF l. The concentration of absent TFs is zero. We assume that all TFs have same concentration ($C_l = C$), as in the analytical calculation.

3. compute the probabilities that a crosstalk state occurs at any given gene, using the thermodynamic model. Other parameters include the energy difference between unbound and cognate state E_a which does not affect the final crosstalk result, and the concentration of the transcription factors, C.

4. compute probability of crosstalk state occurrence for every gene and then obtain the total crosstalk X by summing all these individual contributions of the genes.

5. average over a large number of realizations (we use several hundred realizations for which the average crosstalk has already converged).

6. repeat this procedure (each with multiple realizations) using a different concentration value C each time. Then pick the one that yields the lowest crosstalk value to be $X^*(t)$.

Numerical computation of crosstalk where each TF regulates multiple genes. In an actual gene regulatory network, many TFs regulate multiple genes and many genes are regulated by multiple TFs rather than the on-to-one TF-gene association we considered so far. Specifically, in our data, around 96% of the genes are regulated by more than one TF. To account for that, we obtained the list of genes that are regulated by the *S. cerevisiae* given transcription factors [53].

This computation closely follows the previous procedure. The main difference is in the computation of the similarity measure of genes regulated by multiple cognate TFs. We now have multiple binding site sequences per gene (one for each cognate TF) and, as a result, more binding energies and similarity measures. To take this into account, for each gene we average the similarity measures over its different binding sequences, as follows:

1. for a given gene k, we find all the TFs that regulate it.

2. obtain the consensus sequences of these TFs.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. assume each such consensus sequence represents a potential binding site sequence of gene k (same as in the case of only one TF regulating each gene).

4. assume TF concentrations are proportional to the number of genes they regulate.

5. compute the similarity measure S_{k^i} between each potential binding site sequence i of gene k and all other TFs; this is done in the same way as for one TF regulating one gene using Eq. 10, but now using the weighted TF concentrations.

6. use the mean of the computed S_{k^i} similarity measures taken over the various binding sites of gene k as the effective similarity of that gene.

The rest of the procedure follows the previous calculation, where each TF has only one cognate binding site: i) use the new similarity measure and compute the probability of crosstalk states occurrence, ii) do this for all genes, iii) compute the total crosstalk, iv) average over different random choices of the set of t genes, and v) minimize X with respect to the concentration C to obtain the $X^*(t)$.

Some technicalities and concerns regarding PCM usage When computing the energy matrices using PCMs, certain issues arise that could strongly bias the results if not properly addressed;

- *Inequality of total counts between positions* in PCM data. The sum of counts over all 4 nucleotides in a given PCM should be equal for all positions, but occasionally positions with different total counts are found. As they bias our occurrence statistics (and hence our energy calculation), we used only PCMs in which the total count is equal throughout.

- *Zero counts* in the PCMs. Many PCMs include zero counts for certain nucleotides at specific positions, rendering that element of the energy matrix undefined. Here, we apply a commonly used practice of adding a pseudocount δ to all PCM entries. Following a previous work [13], where various δ values were compared to an information method (where pseudocount is not needed), we take $\delta = 0.1$.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- *Count number sufficiency.* To achieve a reliable estimation of energies in the energy matrix, we only used PCMs with at $p_{\text{counts}} = 5$ count numbers per positions.

Altogether we found 196 TF PCMs, but due to the above concerns, we considered only 23 of them in our calculations.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements We thank U. Alon and Y. Pilpel for raising the question that triggered this study. We thank Nick Barton, Yonatan Friedman, Avi Mayo, Tiago Paixão, Gašper Tkačik and Marcin Zagorski for comments on the manuscript.

References

[1] Michael A. Savageau. Genetic Regulatory Mechanisms and the Ecological Niche of Escherichia coli. *Proceedings of the National Academy of Sciences*, 71(6):2453–2455, June 1974.

[2] M. A. Savageau. Design of molecular control mechanisms and the demand for gene expression. *Proceedings of the National Academy of Sciences*, 74(12):5647–5651, December 1977.

[3] M. A. Savageau. Regulation of differentiated cell-specific functions. *Proceedings of the National Academy of Sciences*, 80(5):1411–1415, March 1983.

[4] Ulrich Gerland and Terence Hwa. Evolutionary selection between alternative modes of gene regulation. *Proceedings of the National Academy of Sciences*, 106(22):8841–8846, June 2009.

[5] Michael A. Savageau. Demand Theory of Gene Regulation. I. Quantitative Development of the Theory. *Genetics*, 149(4):1665–1676, August 1998.

[6] Guy Shinar, Erez Dekel, Tsvi Tlusty, and Uri Alon. Rules for biological regulation based on error minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):3999–4004, March 2006.

[7] Vered Sasson, Irit Shachrai, Anat Bren, Erez Dekel, and Uri Alon. Mode of Regulation and the Insulation of Bacterial Gene Expression. *Molecular Cell*, 46(4):399–407, May 2012.

[8] Arthur L. Koch. The protein burden oflac operon products. *Journal of Molecular Evolution*, 19(6):455–462, November 1983.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[9] C. G. Kurland and Henjiang Dong. Bacterial growth inhibition by overproduction of protein. *Molecular Microbiology*, 21(1):1–4, July 1996.

[10] Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–592, July 2005.

[11] Mahendra Kumar Prajapat, Kirti Jain, Debika Choudhury, Nikhil Raj, and Supreet Saini. Revisiting demand rules for gene regulation. *Molecular BioSystems*, 12(2):421–430, 2016.

[12] Sarah A. Cepeda-Humerez, Georg Rieckh, and Gašper Tkačik. Stochastic Proofreading Mechanism Alleviates Crosstalk in Transcriptional Regulation. *Physical Review Letters*, 115(24):248101, December 2015.

[13] Tamar Friedlander, Roshan Prizak, Călin C. Guet, Nicholas H. Barton, and Gašper Tkačik. Intrinsic limits to gene regulation by global crosstalk. *Nature Communications*, 7:12307, August 2016.

[14] Madeline A. Shea and Gary K. Ackers. The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation. *Journal of Molecular Biology*, 181(2):211–230, January 1985.

[15] P. H. Von Hippel and O. G. Berg. On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences*, 83(6):1608, 1986.

[16] Ulrich Gerland, J. David Moroz, and Terence Hwa. Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, 2002.

[17] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124, April 2005.

[18] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, May 2010.

[19] Michael Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(6):1–21, 2007.

[20] A. Sarai and Y. Takeda. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proceedings of the National Academy of Sciences*, 86(17):6513–6517, September 1989.

[21] Sebastian J. Maerkl and Stephen R. Quake. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, 315(5809):233–237, January 2007.

[22] Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–440, October 2009.

[23] Polly M. Fordyce, Doron Gerber, Danh Tran, Jiashun Zheng, Hao Li, Joseph L. DeRisi, and Stephen R. Quake. *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotechnology*, 28(9):970–975, September 2010.

[24] Ariel Afek, Joshua L. Schipper, John Horton, Raluca Gordân, and David B. Lukatsky. Protein-DNA binding in the absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences*, page 201410569, October 2014.

[25] Shalev Itzkovitz, Tsvi Tlusty, and Uri Alon. Coding limits on the number of transcription factors. *BMC Genomics*, 7(1):239, September 2006.

[26] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, December 2005.

[27] Yufang Wang, Ling Guo, Ido Golding, Edward C. Cox, and N. P. Ong. Quantitative Transcription Factor Binding Kinetics at the Single-Molecule Level. *Biophysical Journal*, 96(2):609–620, January 2009.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Tamar Friedlander, Roshan Prizak, Nicholas H. Barton, and Gašper Tkačik. Evolution of new regulatory functions on biophysically realistic fitness landscapes. *Nature Communications*, 8(1):216, August 2017.

[29] Tsvi Tlusty, Albert Libchaber, and Jean-Pierre Eckmann. Physical Model of the Genotype-to-Phenotype Map of Proteins. *Physical Review X*, 7(2):021037, June 2017.

[30] A. Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, 2008.

[31] O. C. Martin, A. Krzywicki, and M. Zagorski. Drivers of structural features in gene regulatory networks: From biophysical constraints to biological function. *Physics of Life Reviews*, 17:124–158, July 2016.

[32] Tamar Friedlander, Avraham E. Mayo, Tsvi Tlusty, and Uri Alon. Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems. *PLoS ONE*, 8(8), August 2013.

[33] Joshua L. Payne and Andreas Wagner. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science*, 343(6173):875–877, February 2014.

[34] Mark G. Carter, Alexei A. Sharov, Vincent VanBuren, Dawood B. Dudekula, Condie E. Carmack, Charlie Nelson, and Minoru SH Ko. Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biology*, 6:R61, June 2005.

[35] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, February 2014.

[36] Nagarjuna Nagaraj, Jacek R. Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology*, 7(1):548, January 2011.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[37] Han Tao, Christoph Bausch, Craig Richmond, Frederick R. Blattner, and Tyrrell Conway. Functional Genomics: Expression Analysis of Escherichia coli Growing on Minimal and Rich Media. *Journal of Bacteriology*, 181(20):6425–6440, October 1999.

[38] Yan Wei, Jian-Ming Lee, Craig Richmond, Frederick R. Blattner, J. Antoni Rafalski, and Robert A. LaRossa. High-Density Microarray-Mediated Gene Expression Profiling of Escherichia coli. *Journal of Bacteriology*, 183(2):545–556, January 2001.

[39] Sina Ghaemmaghami, Won-Ki Huh, Kiowa Bower, Russell W. Howson, and others. Global analysis of protein expression in yeast. *Nature*, 425(6959):737, 2003.

[40] *Genes IX 9th edition by Lewin, Benjamin published by Jones & Bartlett Publishers Hardcover*. Jones & Bartlett Publishers, March 2007.

[41] Morgan N. Price, Adam M. Deutschbauer, Jeffrey M. Skerker, Kelly M. Wetmore, Troy Ruths, Jordan S. Mar, Jennifer V. Kuehl, Wenjun Shao, and Adam P. Arkin. Indirect and suboptimal control of gene expression is widespread in bacteria. *Molecular Systems Biology*, 9(1):660, January 2013.

[42] Ruth Hershberg and Hanah Margalit. Co-evolution of transcription factors and their targets depends on mode of regulation. *Genome Biology*, 7:R62, 2006.

[43] Irit Shachrai, Alon Zaslaver, Uri Alon, and Erez Dekel. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. *Molecular Cell*, 38(5):758–767, June 2010.

[44] Matthew W. Hahn, Jason E. Stajich, and Gregory A. Wray. The Effects of Selection Against Spurious Transcription Factor Binding Sites. *Molecular Biology and Evolution*, 20(6):901–906, June 2003.

[45] Long Qian and Edo Kussell. Genome-Wide Motif Statistics are Shaped by DNA Binding Proteins over Evolutionary Time Scales. *Physical Review X*, 6(4):041009, October 2016.

[46] Ali Zarrinpar, Sang-Hyun Park, and Wendell A. Lim. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–680, December 2003.

bioRxiv preprint doi: https://doi.org/10.1101/372672; this version posted July 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[47] Jingshan Zhang, Sergei Maslov, and Eugene I. Shakhnovich. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Molecular Systems Biology*, 4(1):210, January 2008.

[48] Arvind Murugan, James Zou, and Michael P. Brenner. Undesired usage and the robust self-assembly of heterogeneous structures. *Nature communications*, 6, 2015.

[49] Muir J. Morrison and Justin B. Kinney. Modeling multi-particle complexes in stochastic chemical systems. *arXiv:1603.07369 [cond-mat, physics:physics, q-bio]*, March 2016. arXiv: 1603.07369.

[50] Lagator Mato, Paixão Tiago, Nicholas H. Barton, Jonathan P. Bollback, and Călin C. Guet. On the mechanistic nature of epistasis in a canonical cis-regulatory element. *eLife; Cambridge*, 6, 2017.

[51] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, 11(12):4241–4257, December 2000.

[52] Aaron T. Spivak and Gary D. Stormo. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Research*, 40(D1):D162–D168, January 2012.

[53] Saccharomyces Genome Database | SGD.