PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

The relation between crosstalk and gene regulation form revisited

Rok Grah[1], **Tamar Friedlander**[2]*

1 Institute of Science and Technology Austria, Klosterneuburg, Austria, **2** The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot, Israel

* tamar.friedlander@mail.huji.ac.il

OPEN ACCESS

Citation: Grah R, Friedlander T (2020) The relation between crosstalk and gene regulation form revisited. PLoS Comput Biol 16(2): e1007642. https://doi.org/10.1371/journal.pcbi.1007642

Editor: Reka Albert, Pennsylvania State University, UNITED STATES

Received: March 20, 2019

Accepted: January 8, 2020

Published: February 25, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: RG is a recipient of a DOC Fellowship (Doctoral Fellowship Programme of the Austrian Academy of Sciences, https://stipendien.oeaw.ac.at/en/stipendien/doc/) of the Austrian Academy of Sciences at the Institute of Science and Technology Austria (grant number 25006). TF acknowledges financial support from the Hebrew University of Jerusalem. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Genes differ in the frequency at which they are expressed and in the form of regulation used to control their activity. In particular, positive or negative regulation can lead to activation of a gene in response to an external signal. Previous works proposed that the form of regulation of a gene correlates with its frequency of usage: positive regulation when the gene is frequently expressed and negative regulation when infrequently expressed. Such network design means that, in the absence of their regulators, the genes are found in their least required activity state, hence regulatory intervention is often necessary. Due to the multitude of genes and regulators, spurious binding and unbinding events, called "crosstalk", could occur. To determine how the form of regulation affects the global crosstalk in the network, we used a mathematical model that includes multiple regulators and multiple target genes. We found that crosstalk depends non-monotonically on the availability of regulators. Our analysis showed that excess use of regulation entailed by the formerly suggested network design caused high crosstalk levels in a large part of the parameter space. We therefore considered the opposite 'idle' design, where the default unregulated state of genes is their frequently required activity state. We found, that 'idle' design minimized the use of regulation and thus minimized crosstalk. In addition, we estimated global crosstalk of *S. cerevisiae* using transcription factors binding data. We demonstrated that even partial network data could suffice to estimate its global crosstalk, suggesting its applicability to additional organisms. We found that *S. cerevisiae* estimated crosstalk is lower than that of a random network, suggesting that natural selection reduces crosstalk. In summary, our study highlights a new type of protein production cost which is typically overlooked: that of regulatory interference caused by the presence of excess regulators in the cell. It demonstrates the importance of whole-network descriptions, which could show effects missed by single-gene models.

Author summary

Genes differ in the frequency at which they are expressed and in the form of regulation used to control their activity. The basic level of regulation is mediated by different types of

Competing interests: The authors have declared that no competing interests exist.

DNA-binding proteins, where each type regulates particular gene(s). We distinguish between two basic forms of regulation: positive—if a gene is activated by the binding of its regulatory protein, and negative—if it is active unless bound by its regulatory protein. Due to the multitude of genes and regulators, spurious binding and unbinding events, called "crosstalk", could occur. How does the form of regulation, positive or negative, affect the extent of regulatory crosstalk? To address this question, we used a mathematical model integrating many genes and many regulators. As intuition suggests, we found that in most of the parameter space, crosstalk increased with the availability of regulators. We propose, that crosstalk is usually reduced when networks are designed such that minimal regulation is needed, which we call the 'idle' design. In other words: a frequently needed gene will use negative regulation and conversely, a scarcely needed gene will employ positive regulation. In both cases, the requirement for the regulators is minimized. In addition, we demonstrate how crosstalk can be calculated from available datasets and discuss the technical challenges in such calculation, specifically data incompleteness.

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 regulated using either mode, 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 was addressed using different approaches. The seminal work of Michael Savageau [1–3] 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. Savageau argued that the intensity of selection depends on the extent to which the regulatory construct is used (later called the "use it or lose it" principle [4]). When infrequently used (as in activator regulating a low-demand or a repressor regulating a high-demand gene), selection to preserve is weak, rendering it unlikely 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 [4].

Recently, a comprehensive survey of regulatory topologies in *E. coli* and *B. subtilis*, found agreement between the experimentally observed topologies and their satisfaction of dynamic constraints, as verified in simulations. The authors found exceptions to the Savageau demand rule and proposed that evolutionary processes randomly pick a regulatory topology out of the many possible ones meeting the organism physiological constraints [6].

An alternative reasoning for the observed correlation between a gene's demand and its form of regulation was proposed using a biophysical, rather than evolutionary argument [7, 8]. If a high-demand gene is regulated by an activator and a low-demand gene is regulated by a repressor, their regulatory binding sites are mostly occupied and protected from spurious binding of foreign regulators that could interfere with the gene's regulatory state. However, if this reasoning applies not just to one gene, but to many of them, it would also entail

extravagant use of regulators [6]. This would place heavy demands on protein expression systems, associated with reduced growth rate [9–13].

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, regardless of the full regulatory network. It remains unanswered whether the choice of positive or negative regulation for a gene with low- or high-demand could have additional costs for the entire network. Specifically, transcription factors are known to have limited specificity and bind a variety of DNA targets, besides their cognate binding sites [14–19]. The probability of such binding events naturally depends on their concentrations [20, 21]. Here, we revisit the argument that the Savageau demand rule minimizes transcriptional crosstalk, by accounting for crosstalk of multiple genes simultaneously, rather than the single-gene crosstalk considered earlier.

We use a mathematical global crosstalk model [22], which was built upon the well-established thermodynamic model of gene regulation to calculate transcription factor (TF)-DNA interactions [20, 21, 23–27]. 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 [22]. In contrast, the *global* crosstalk cannot be reduced below a certain threshold. Here, we analyze global crosstalk levels under different regulatory strategies: either positive or negative regulation. We compare two extreme designs: 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' design, in which a high (low)-demand gene is always regulated by a repressor (activator). We find that the 'busy' design maximizes regulator usage, whereas the 'idle' one minimizes it. We analyze the dependence of global crosstalk on the abundance of regulatory proteins in the cellular environment and find the exact conditions under which either 'idle' or 'busy' design minimizes crosstalk. We conclude that under most biologically plausible parameter values, the 'idle' design 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 designs, i.e., 'busy' and 'idle' and then construct numerical simulations of a more general asymmetric gene usage model, that are in agreement with the analytical result. Lastly, we discuss the challenges in crosstalk calculation for real gene regulatory networks, in particular, the possible effect of data incompleteness, and show an example using *S. cerevisiae* TF data.

Results

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

We begin by introducing and analyzing a basic model with a simple form of gene regulation, assuming that each gene is regulated by a single transcription factor. We also assume identical properties for all genes and all transcription factors. Later we relax some of these simplifying assumptions and consider additional more complex gene regulatory architectures. We summarize these model variants in the main text, and their full descriptions can be found in S1 Text. 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 $0 \leq p \leq 1$ 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 the M genes. We assume however that all these subsets include an equal q proportion of genes $0 \leq q \leq 1$ that is needed to be active. The remaining $1 - q$ proportion should be 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, is available to bind and regulate these genes. However, TFs often have limited specificity to their DNA targets and can occasionally bind slightly different sequences, albeit with lower probability [16, 17, 19, 28–30].

We define 'crosstalk' as the average fraction of genes found in any erroneous regulatory state: a gene that should be activated (repressed) but is not, because its cognate TF fails to bind or because its binding site which should remain unoccupied is bound by a non-cognate TF and also events of activation (repression) in response to a non-cognate signal (or in a wrong dynamic range) because a non-cognate activator (repressor) binds instead of the cognate one—see summary in Fig 1C. To quantitate the probability of these events, we use the thermodynamic model of gene regulation [20, 21, 23, 24, 31]. Importantly, this model assumes that gene activity is proportional to the equilibrium binding probability of its transcription factor to its regulatory binding site. Hence, we use a quasi-static, rather than kinetic, description where we assume that the system switches between different states of equilibrium. A mathematical model for crosstalk for the special case in which all TFs are activators ($p = 0$) was derived and analyzed in our previous work [22]. Here, we analyze a more general model with a combination of activators and repressors. The reader can find the details of both models in S1 Text.

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 1 and Fig 1B).

The probability that a particular gene i is in the x_{bound} or x_{unbound} crosstalk states, depends on the concentration 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 $j \neq i$ [22]:

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

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. Eq (1) can also be used for general TF concentrations, as observed in experiments. We demonstrate this calculation in S1 Text (Section 8.4). We found that allowing different concentrations for activators and repressors does not reduce crosstalk below this equal concentration scheme (S1 Text). We also assume full symmetry between binding sites i, such that $S_i = S \; \forall i$. A

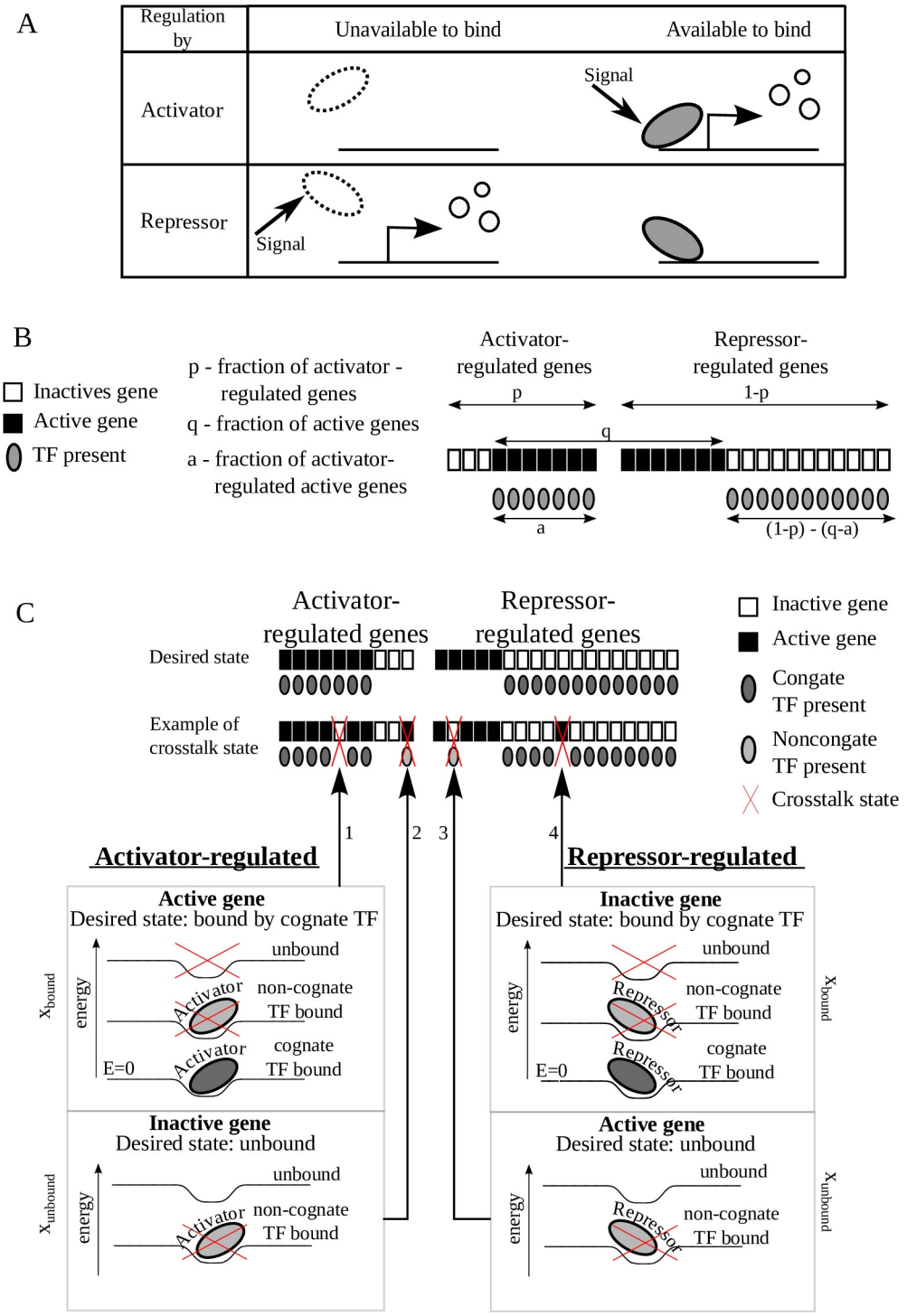

Fig 1. Gene regulation can employ different combinations of activators and repressors to implement the same gene expression pattern. (A) A signal can cause gene activation by either positive (first row) or negative (second row) control. **(B)** We consider a total of M genes in a cell, of which a fraction $0 \leq p \leq 1$ is regulated by activators, and the remaining $1 - p$ 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, $a \leq q, p$ 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 at different levels of energy depending on its 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). The upper panel shows the crosstalk-free 'desired state' (first row), where each TF binds its cognate target. Below (second row), four different possibilities in which binding of a TF to non-cognate binding sites or failure to bind 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 marked by red crosses.

https://doi.org/10.1371/journal.pcbi.1007642.g001

numerical analysis of a more general case with non-uniform S_i values can be found in Fig B in S1 Text. The value of S can be either estimated using binding site data (see below) or analytically calculated under 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 [20, 22, 24]:

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

$$x_{\text{unbound}} = \frac{\sum_{j\neq i} C_j e^{-\epsilon d_{ij}}}{e^{-E_a} + \sum_{j\neq i} C_j e^{-\epsilon d_{ij}}} = \frac{cs}{e^{-E_a} + cs}. \tag{3}$$

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.

Table 1. We distinguish 4 sets of genes according to their state of activity (active/inactive) and form of regulation (activation/repression).

Activity	Regulated by	Proportion of genes using this regulatory strategy
active	activator	a, where $a \leq q, p$
active	repressor	$q - a$
inactive	activator	$p - a$
inactive	repressor	$(1 - p) - q + a$

https://doi.org/10.1371/journal.pcbi.1007642.t001

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 over different choices of $q - a$ out of the $(1 - p)$ repressor-regulated proportions. The weighted sum over these four types of contributions provides the average total crosstalk, X, of the whole system:

$$\begin{aligned} X &= \overbrace{a \cdot x_{\text{bound}} + (p-a) \cdot x_{\text{unbound}}}^{\text{Contribution of activator-regulated genes}} + \overbrace{(q-a) \cdot x_{\text{unbound}} + (1-p-q+a) \cdot x_{\text{bound}}}^{\text{Contribution of repressor-regulated genes}} \\ &= t \cdot x_{\text{bound}} + (1-t) \cdot x_{\text{unbound}}. \end{aligned} \tag{4}$$

As Eq (4) 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 [22], 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} (Eqs (2) and (3) into Eq (4)). We then minimize X with respect to the total TF concentration, c. Such minimization is possible because global crosstalk balances between some binding sites that should be bound and others that should be unbound. For the former, higher c increases their chance to be bound by their cognate TFs and thus reduces crosstalk. For the latter, their cognate TF is absent and thus higher c increases their chance to be bound by foreign TFs, namely increases crosstalk. We then obtain the expression for minimal crosstalk:

$$X^*(t,s) = t(-s(1-t) + 2\sqrt{s(1-t)}). \tag{5}$$

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, the larger 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 S1 Text that $t^*(s) \geq 2/3$), X^* increases with t. Intuitively, the number of available TF species 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 since the optimal concentration $c^*(s, t)$ is tailored specifically for each t value. That is because the relative weight of binding sites that should be bound vs. those that should be unbound, shifts with t. High TF usage though always comes at the cost of an exponential increase in the optimal TF concentration, c^*, (Eq. S4), where for high s values, c^* diverges to infinity $c^* \to \infty$ (see Fig 2B). We discuss below the biological relevance of the high t regime. We derived this model for the simple regulatory network shown in Fig 1C. Eqs (2)–(5) can be analogously derived for more complex network architectures, as we demonstrate in S1 Text (Section 10).

Mode of regulation affects global crosstalk because it affects TF usage

A particular 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 (5), 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

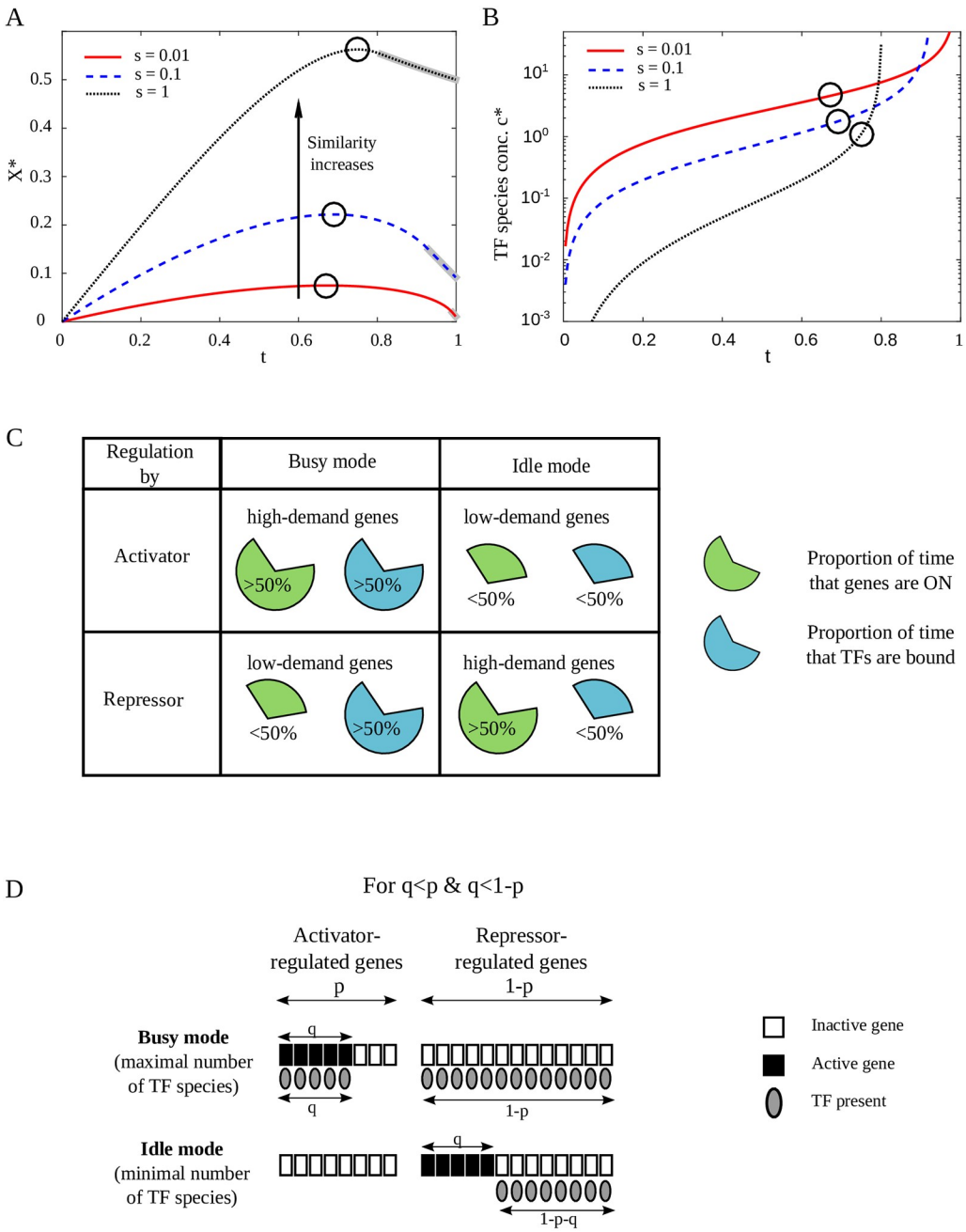

Fig 2. Crosstalk depends on the fraction of available TFs, which varies between regulatory designs. (A) We illustrate minimal crosstalk, X^*, vs. t, the fraction of available TFs, for different values of similarity, s. In most of the parameter regime (for $t < t^*$, $t^* \geq 2/3$), minimal crosstalk, 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 gray-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 design'). 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 for a small fraction of the time ('idle design'). **(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') design, as many active genes as possible are assigned positive (negative)

regulation and as many inactive genes as possible are assigned negative (positive) regulation. The scheme shows an example with the proportion of active genes q, the proportion of activator-regulated genes p and the proportion of repressor-regulated genes $(1 - p)$ such that $q \leq p, 1 - p$. Other combinations are shown in Fig E in S1 Text.

https://doi.org/10.1371/journal.pcbi.1007642.g002

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, thereby covering 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 S1 Text for an example). Specifically, we focus on the two extreme gene-regulator associations, which we call the 'busy' and 'idle' network designs. The 'busy' design means that gene regulation is operative most of the time. It is implied by the "Savageau demand rule" [2], because the gene's default state of activity is not its commonly needed state. Under the opposite 'idle' design, 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 designs are also possible.

To represent the 'busy' design, we associate as much of 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 is regulated by repressors. Thus the fraction of activator-regulated active genes is $a = \min(p, q)$. Conversely, under the 'idle' design, we associate as much of 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 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:

$$t_{\text{busy}} = 1 - |p - q|, \tag{6}$$

$$t_{\text{idle}} = |1 - p - q|. \tag{7}$$

In Fig 2D, we illustrate regulation following these two extreme designs. The TF assignments defined in Eqs (6) and (7) 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 $t_{\text{idle}} \leq t \leq t_{\text{busy}}$ (see S1 Text for formal proof). In Fig 3A, we illustrate the difference in the fraction of available TFs between the two extreme designs $\Delta t = t_{\text{busy}} - t_{\text{idle}} = 1 - |p - q| - |1 - p - q| > 0$, demonstrating that the 'busy' design always requires more regulators than the 'idle' design (see S1 Text).

Using Eq (5), we obtain exact expressions for X^* under these extreme designs (see S1 Text). In Fig 3B, we show $\Delta X^* = X^*_{\text{idle}} - X^*_{\text{busy}}$, the difference in minimal crosstalk X^* between the two extreme designs, for all (p, q) combinations. We find that the 'idle' design yields less crosstalk in a large part of this parameter space. The 'busy' design still involves less crosstalk for parameter combinations centered around the diagonal $p = q$, whereas the 'idle' design always performs best on the anti-diagonal $1 - p = q$. This is due to the fact that 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' design most advantageous.

In the region in which the 'busy' design yields the lowest crosstalk, this comes at the cost of using a larger fraction of existing TF species, as depicted in Fig 3C. The 'idle' design, in contrast, requires a much smaller fraction of TF species. Furthermore, the two designs differ not

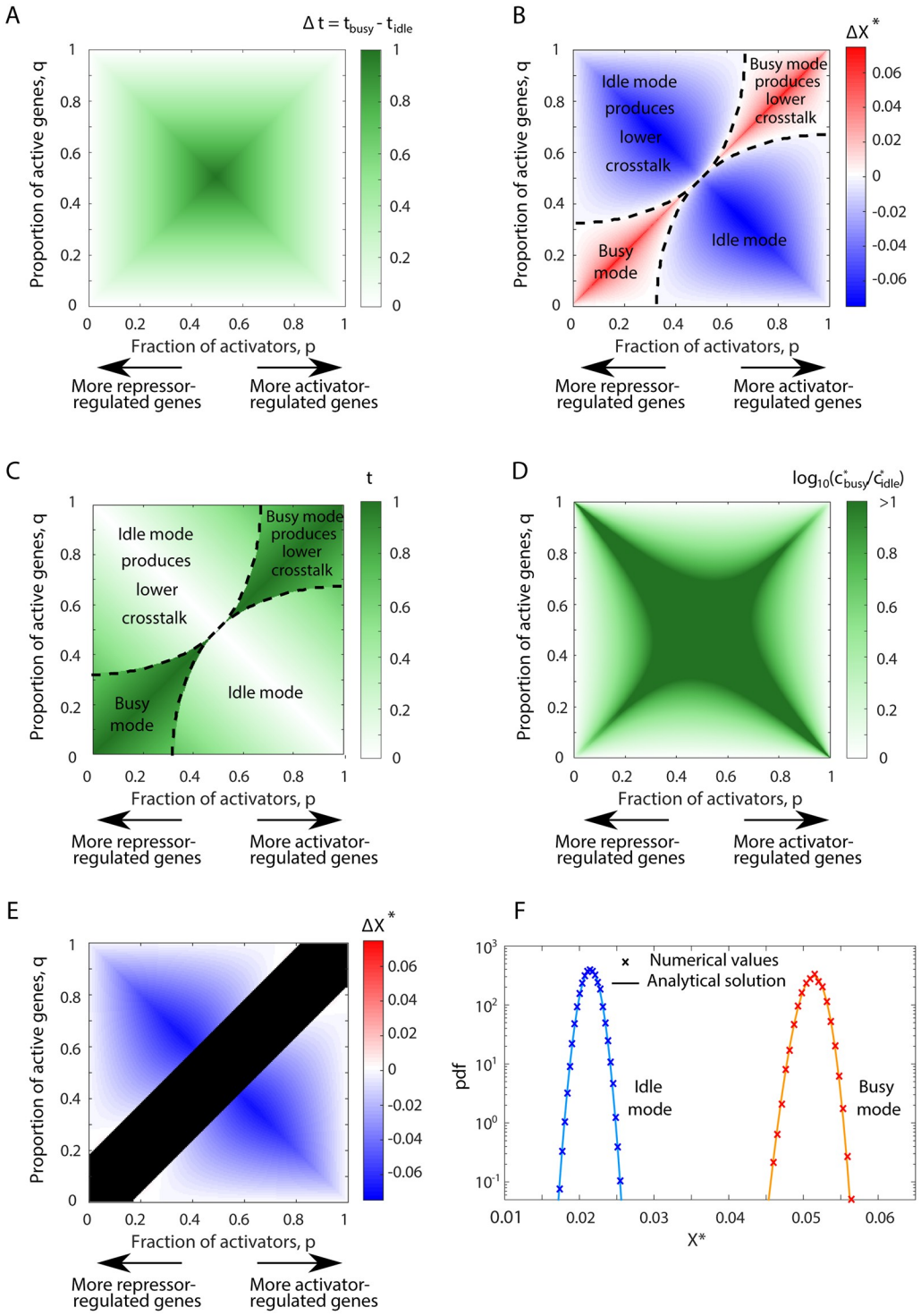

Fig 3. 'Idle' design yields lower crosstalk than the 'busy' in a large part of the parameter regime. (A) The 'busy' design always requires more TFs compared to the 'idle' design. Here we illustrate Δt, the difference in the fraction of TFs in use between the two designs for different values of p and q (shown in color scale). **(B)** The difference in minimal total crosstalk ($\Delta X = X^*_{\text{idle}} - X^*_{\text{busy}}$) between 'idle' and 'busy' designs, shown in color scale, as a function of p and q. In a large part of the parameter regime (colored blue), lower crosstalk is achieved by the 'idle' design. The 'busy' design is most beneficial on the diagonal $p = q$ (red region), but this requires use of all TFs and comes at the cost of an enormously high TF concentration. The 'idle' design is most beneficial around the anti-diagonal $q = 1 - p$, where regulation can proceed with no TFs at all and crosstalk is close to zero. **(C)** Fraction of TFs in use (shown in color scale) when the design

providing minimal crosstalk ('idle' or 'busy' as in (B)) is used, as a function of p and q. Black dashed lines mark the borders between the regions where 'busy' or 'idle' designs provide lower crosstalk. While 'idle' design mostly requires a minority (< 50%) of the TFs, the 'busy' design 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 design $c^*_{\text{busy}}/c^*_{\text{idle}}$. 'Busy' design always requires higher TF concentrations. **(E)** For higher similarity s between binding sites, parts of the parameter space fall 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 designs for $s = 1$. Black areas denote the anomalous regime. Importantly, the region where the 'busy' design was beneficial for low s (see (B)) falls into this anomalous regime. **(F)** Analytical solution of the stochastic model for the distribution of crosstalk values, is in excellent agreement with stochastic 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' design and $\gamma_i = \gamma = 0.92$ for 'busy' design, with $2 \cdot 10^6$ realizations.

https://doi.org/10.1371/journal.pcbi.1007642.g003

only in the fraction of TFs needed but also in their concentrations. To achieve the lower bound, the 'busy' design always requires a higher total TF concentration, c^* (Fig 3D).

The explanation for the alternating crosstalk advantage between the two extreme designs 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' design will yield lower crosstalk, whereas if $t^* < t_{\text{idle}} < t_{\text{busy}}$, 'busy' will be more advantageous (see S1 Text for more details). While 'idle' and 'busy' represent the two extremes, a continuum of regulatory designs 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 S1 Text).

We previously found that for some parameter combinations of similarity, s, and fraction of active genes, q, the mathematical expression for X^* (Eq (5)) has no biological relevance [22]. 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. Another biologically irrelevant regime occurs for high TF usage $t > t_{\max}$ (see SI of [22]). Then the concentration needed to obtain minimal crosstalk formally diverges to infinity $c^* \to \infty$. These biologically implausible regimes put an upper bound to the total number of genes that an organism can effectively regulate [22, 32]. 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' design 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' design yields lower crosstalk in the entire biologically relevant parameter space—see S1 Text and Figs F, G in S1 Text.

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 [33, 34]. 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, $i = 1 \ldots M$, 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, where $\gamma_i = D_i$ if regulation is positive and $\gamma_i = 1 - D_i$ if it is negative. We then obtain exact solutions for the distributions of t and X^* (Eqs. S14-S15 in S1 Text). In Fig 3F, we illustrate the X^* distributions for two values of t, representative of the two extreme designs. We find excellent agreement between this analytical solution and stochastic simulation results. 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 S1 Text.

Data-based crosstalk calculation

Similarity and crosstalk, considered in our analytical model, can be estimated from bioinformatic data. As direct thorough measurements of TF binding preferences are available for only a few TFs [16, 29, 30], we use statistical estimates based on multiple binding sites to which a particular TF binds (PCM) to determine its binding energetics to various sequences. Specifically, we use data of 23 *S. cerevisiae* transcription factors collected from the scerTF database [35, 36]. 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 and then numerically calculate crosstalk.

In our theoretical model, we made 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 that all TFs regulate an equal number of genes (a single gene per TF, in the basic model). The availability of TF binding data allows us to relax these assumptions, and consider variation in binding energies and promiscuity among TFs, as well as the actual unequal energetic contributions of the different positions in each binding site.

For simplicity, we still assume equal concentrations for all available TFs and calculate a lower bound on crosstalk if concentrations are optimized. In S1 Text (Section 8.4) we demonstrate how crosstalk calculation can be implemented for general TF concentration values and show an example using experimentally measured concentrations [37]. Due to paucity of data on epistatic effects between distinct binding site positions, we still assume additivity in the energetic contributions of different positions in the sequence. The latter assumption is considered reasonable for up to 3-4 bp substitutions [16].

Similarity values vary between genes even within the same organism. We begin by numerically calculating the similarity s_i between consensus binding sequences of different transcription factors (see Methods). In Fig 4A, we show the distribution of similarity values of genes associated with 23 *S. cerevisiae* transcription factors (top). We find a broad distribution of s_i values spanning over 5 orders of magnitude, where its median is around $10^{-4} - 10^{-3}$. This finding is in marked contrast to the full symmetry and equal s_i values for all TFs assumed in our analytical solution.

While we find that s_i values are very variable, the largest contributions to global crosstalk are made by the few most promiscuous TFs (those with high s_i values). In the following, we fit an effective similarity value that would best capture the numerically calculated crosstalk values, had all TFs had uniform s_i values, as in the mathematical model (denoted by red arrow in Fig 4A). In this example, we find that $s_{\text{effective}}$ is almost equal to the median s_i value (black arrow there).

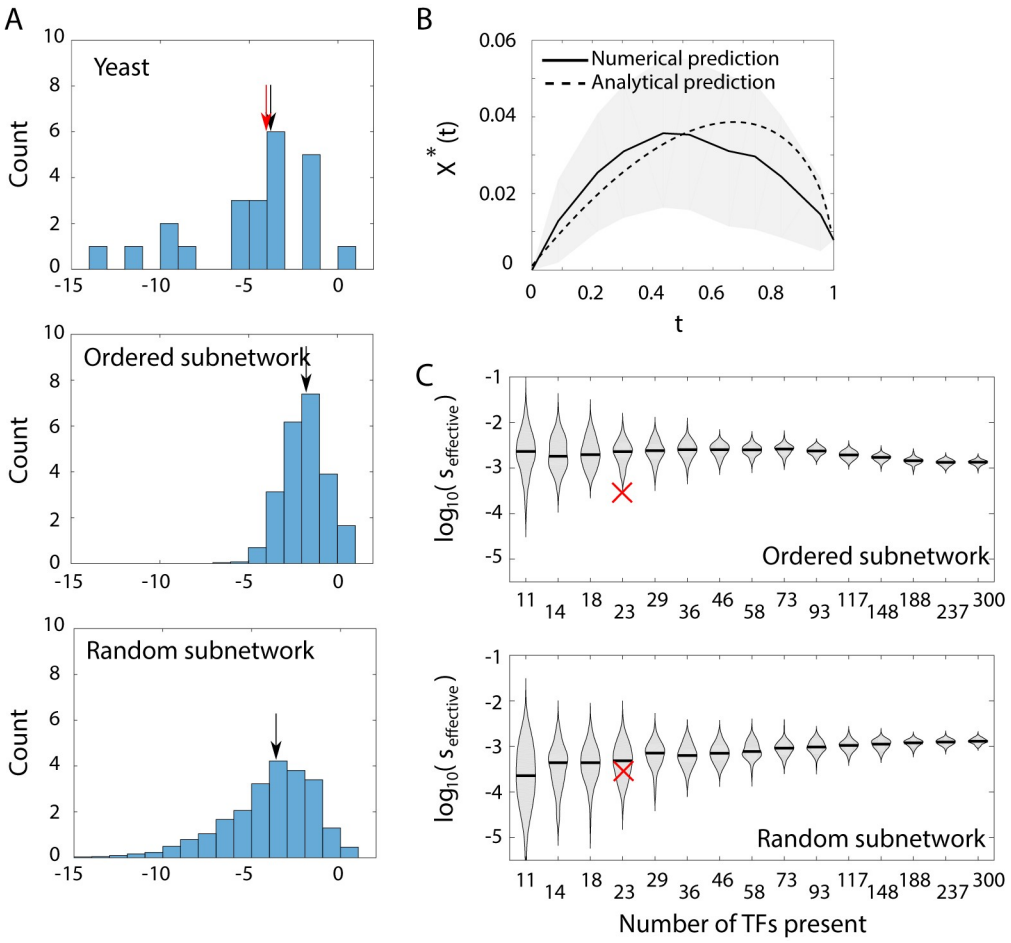

Fig 4. Data-based crosstalk estimates. (A) Inter-TF similarity values of *S. cerevisiae* TFs (top), and of synthetic data (middle and bottom) exhibit broad distributions spanning a few orders of magnitude. The distribution median values are marked by black arrows. The red arrow in the yeast data represents $s_{effective}$ of the yeast data, and nearly overlaps with the distribution median. Synthetic data were created by randomly drawing PCMs representing all TFs of an artificial network. Then, sub-networks of 23 TFs were sampled by either taking the 23 most promiscuous TFs (middle) or randomly choosing them (bottom). The figures show similarity distributions amongst TFs in these artificial networks, averaged over 100 repeated draws. s_i values here are with respect to all TFs in the network, regardless of their (un)availability. **(B)** Numerical prediction of minimal global crosstalk depending on TF availability t for *S. cerevisiae* (solid line) compared to an analytical prediction based on a single $s_{effective}$ value common to all genes (dashed line). This effective similarity value was chosen to provide the best fit to the numerical curve. The curves represent estimation of crosstalk for the network of all 2126 *S. cerevisiae* downstream genes regulated by the 23 TFs, for which we have PCMs. The numerical curve represents the mean over 10^3 realizations for each t value, where the exact subset of available TFs was randomly drawn. The surrounding gray shadings show ±1 standard deviation around the mean. The discrepancy between numerical and analytical calculations is attributed to the broad distribution of s_i values for the numerical calculation, whereas the analytical calculation assumes a uniform s_i value for all TFs. **(C)** Violin plots of $s_{effective}$ for different subnetwork sizes for *ordered* and *random* subnetworks. Ordered subnetworks are the subsets of TFs having highest similarity s_i with respect to the whole network. Random subnetworks include a random subset of the full network TFs. For each subnetwork, we numerically calculated crosstalk and fitted the $s_{effective}$ which would best capture the crosstalk function if all TFs had a uniform s value. The violin plots represent distributions of effective similarity values from 100 different randomly drawn subnetworks, each coming from an independently drawn full network of 300 TFs. The red x represents the $s_{effective}$ value of the 23 yeast TFs (same value as the red arrow in A). For details on the numerical calculations of similarities and crosstalk, see Methods. All violin plots exhibit broad $s_{effective}$ distributions which are broadest for the smallest subnetworks, as expected. For "ordered" subnetworks, the median $s_{effective}$ value is high for the small subnetworks (which were chosen to contain the most promiscuous TFs) and then slightly decreases for bigger subnetworks. For random subnetworks, the trend is opposite.

https://doi.org/10.1371/journal.pcbi.1007642.g004

Numerical crosstalk calculation: Incorporation of a complex TF-gene interaction network. In the analytical model, 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 expand our dataset to include all the 2126 genes [38] regulated by the 23 *S. cerevisiae* TFs for which we have PCMs and considered all possible TF-gene interactions in this set. Notably, there is high variability in the number of genes regulated by each TF. For different values of t (proportion of available TFs), we randomly choose a subset of TFs to be available and accordingly compute the crosstalk probabilities for all genes, accounting for all possible TF-binding site (BS) combinations. We repeat this procedure for 20 different t values, with 100 independent draws of available TFs for each. In the crosstalk calculation, we assume that all available TFs have equal concentrations. In contrast to the analytical calculation, where we included crosstalk contributions from all TFs, here, only binding states associated with transcription factors that are chosen to be available, are considered. In the analytical model we assumed full symmetry between all TFs and all binding sites. Hence a single similarity value s was sufficient. In contrast, in a real network, we obtain a variety of similarity values (Fig 4A). As each TF regulates multiple binding sites, we now calculate similarity between the consensus sequences of the different TFs, and refer to similarity between TFs, rather than similarity between binding sites. In order to compare similarity values of different networks, we fit the numerically calculated crosstalk with the analytical model, where a single $s_{\text{effective}}$ value is used for all TFs. Fig 4B shows both the numerically calculated crosstalk and the analytically predicted one (using $s_{\text{effective}}$) for this more complex interaction network (solid and dashed lines, correspondingly). The gray shading represents ±1 standard deviation around the mean value of the numerically calculated crosstalk.

Data incompleteness could affect crosstalk estimates. Global crosstalk accounts for the combined effects of all of the organism's TFs and binding sites. Unfortunately, data of TF binding preferences is incomplete. Moreover, the accuracy of PCMs depends on the number of known binding sites associated with the TF of interest. Due to these technical limitations, we focused on only 23 *S. cerevisiae* TFs for which > 5 binding sites (per TF) are known. However, this small subset of TFs regulates one third (!) of the yeast genes. Motivated by that, we ask how representative is a crosstalk estimation of the entire network based on this small TF subset. In other words, what fraction of the TFs (or genes they regulate) would suffice to reliably estimate the global network crosstalk. This crosstalk estimation problem is further complicated by the diversity of s_i values we find among TFs. To generally address these questions, we simulate synthetic gene regulatory networks, each integrating 300 TFs. We simulate the binding preferences of these TFs using the PCM statistics of the 23 yeast TFs (see Methods). We then sample subnetworks of different sizes from these full networks and numerically calculate crosstalk for each subnetwork (Methods).

We sample the full networks in two manners: we either randomly choose a subset of TFs ("random subnetworks") or deterministically select the TFs showing the highest similarity with respect to the full network ("ordered subnetworks"). The latter choice is motivated by the prior information that the few yeast TFs for which we have reliable data, are not a random subset, but rather the subset that has the largest number of binding sites. This choice is then a worst-case estimate of global crosstalk. To compare different networks on an equal basis, we estimate the effective similarity $s_{\text{effective}}$ fitted for each subnetwork. Fig 4C shows the distributions and medians of $s_{\text{effective}}$ values obtained, as a function of the subnetwork size. Each distribution is based on independent draws of 100 full networks. From each full network, we sample one random and one ordered subnetwork of each size.

We find, that small-size "ordered" subnetworks exhibit higher median $s_{\text{effective}}$ values but narrower distributions than the "random" subnetworks, as expected. Both "ordered" and "random" subnetworks converge to the same $s_{\text{effective}}$ value for the full network (of size 300). The $s_{\text{effective}}$ distribution for the full size represents variation between various full networks of same size, which is significantly smaller than the variation due to limited sampling, observed for the smaller networks. As the "ordered" subnetworks deliberately include the most promiscuous TFs, their $s_{\text{effective}}$ is an over-estimate of the full network measure. In contrast, we find, that $s_{\text{effective}}$ estimated for random subnetworks is an under-estimate of the full network $s_{\text{effective}}$. In our synthetic data, we allowed for binding site length variation among TFs (the PCM dimension). Interestingly, we find positive correlation between the TF's promiscuity s_i and its consensus binding site length. An opposite effect is found for the length of DNA binding sites (see Fig K in S1 Text).

Considering the sufficiency of the sample size, for an "ordered" subnetwork, a sample of $\sim$ 50 (out of 300) TFs provides variation close to the full network measure, whereas for "random" subnetworks, a larger sample size of around $\sim$ 100 TFs (out of 300) is needed. Either way, we conclude that a global crosstalk estimate is possible with only a subset of the network TFs. We compare our calculated s_i values of yeast data (red cross) to the estimated $s_{\text{effective}}$ distributions of this subnetwork size. Interestingly, the yeast estimated crosstalk value falls below the median value for both "random" and "ordered" sampling approaches. This may imply that selection to reduce crosstalk is at work, yielding similarity values which are lower than what one would expect at random [39].

Discussion

We studied the susceptibility of different gene regulatory networks to transcriptional crosstalk. We found a lower bound on crosstalk $X^* = X^*(t, s)$, which is fully determined by two macroscopic "thermodynamic-like" variables, regardless of other microscopic details of the network. These are the fraction of available TF species, t, and the average similarity between distinct binding site sequences, s. This emergent simplification enabled us to analyze crosstalk for classes of gene regulatory networks, regardless of other network details. We showed that different network designs may vary in t, the TF usage they require, and hence differ in the crosstalk levels they incur, even if they have the same gene activity pattern. We analyzed two extremes: a 'busy' design, which maximizes the use of regulators and is equivalent to the previously proposed Savageau demand rule [1] and the opposite 'idle' design, that minimizes the use of regulators. Interestingly, crosstalk is minimized by either of these extremes, and not by any hybrid design. We found that, in a large part of the parameter regime, crosstalk increased with t, and consequently minimized by the 'idle' design. In the remaining part, crosstalk was minimized by the 'busy' design, but came at a cost of a much higher TF concentration requirement. Our basic analysis refers to a simple network architecture. We exemplify in S1 Text (Section 10) how the crosstalk expressions Eqs (2)–(5) can be generalized to describe more complex regulatory architectures. We also studied a stochastic gene activation variant of the model, where the number of active genes can fluctuate. We found that it is well-approximated by the deterministic activation model, because the distributions of TF availability and minimal crosstalk are typically very narrow and centered around their mean value.

Where are real organisms located in the (t, s) parameter space? Reports of the number of co-expressed genes greatly vary between organisms and depend on growth conditions. For example: $\sim$ 10,000 different genes were reported to be co-expressed in a mouse cell ($<$ 50% of total) [40, 41], 10,000-12,000 ($<$ 50%) genes were estimated to be co-expressed in human HeLa cells

[42], 3300-3500 out of 4290 genes (76%-82%) were co-expressed in *E. coli* during exponential growth [43, 44] and 75%-80% of the genes were co-expressed in *S. cerevisiae* [37, 45].

Values of similarity between distinct TF binding sites, vary not only between organisms, but also between modules and distinct genes within the same organism (see Fig 4). We estimated s_i and the resultant minimal crosstalk values for 23 *S. cerevisiae* TFs using PCM data. We found an extremely broad distribution of single-TF s_i values spanning > 5 orders of magnitude, with a median between $10^{-4} - 10^{-3}$. Global crosstalk, however, is determined by the few high-similarity TFs. To bridge the gap between the high diversity of s_i in real networks and our uniform s analytical solution, we fitted a single $s_{\text{effective}}$ value which would best capture the numerically calculated network crosstalk. For the yeast data, we found that this $s_{\text{effective}}$ is very close to the distribution median. Using our estimates for s and t, we estimated minimal crosstalk X^* for this subnetwork of *S. cerevisiae* to be in the range 0.03-0.04 (see Fig 4B), if 30%-80% of the TFs are present. Our analysis showed that, for relatively low s values, as we found for yeast, there was a regime in the parameter space in which 'busy' yields the lowest crosstalk. The choice of network design that minimizes crosstalk ('busy'/'idle') depends on the proportion of co-activated genes and on the proportion of activators. For organisms with high s values, the regime in which 'busy' is beneficial is actually anomalous, and hence biologically irrelevant. Such higher s is expected for organisms with shorter binding sites.

Binding site data is often incomplete. To assess the validity of whole-network crosstalk estimation based on a small subset of TFs, we constructed synthetic gene regulatory networks, sampled some subnetworks and then compared the s estimation of full and partial networks. In the *S. cerevisiae* case, we found that a full network crosstalk estimate is possible with binding information of only 16%-33% of the TFs.

Here, we used a symmetric and admittedly simplified gene regulatory network model. Our analysis determined a lower bound for crosstalk, assuming that TF concentrations are accurately tuned. In reality, TFs are not necessarily expressed and degraded at a precise time [46] and crosstalk is thus expected to be higher. In S1 Text (Section 8.4) we demonstrate crosstalk calculation with general TF concentrations, obtained in experiments. Relaxation of other simplifying assumptions made in our analytical model opens new research avenues for future work. Most importantly, we assumed uniform similarity values of all TFs and all BS, whereas *S. cerevisiae* data analysis showed diversity in TF properties. In principle, a distribution of s values can be incorporated into the model, but would significantly complicate averaging over different sets of active genes (but see a simple example in S1 Text). Other simplifications include the averaging over gene sets of same-size as representatives of different environmental conditions, whereas, in reality, the number of expressed genes could vary between environments (e.g., growth media [43]). We averaged over all possible choices of active genes, although only some of these activity combinations occur naturally. We also assumed that every gene has a regulator, and vice versa, although this is not always the case. Hershberg and co-workers found an imbalance between genes and regulators, where orphan repressors with no genes and orphan genes with no activators, transiently exist, and could also contribute to crosstalk [47]. Relaxation of these assumptions would require a more comprehensive characterization of gene regulatory networks and co-expression patterns than is known to date.

Our study addressed a typically overlooked cost of protein production: that of regulatory interference caused by excess regulatory proteins in the dense cellular medium. This cost is distinct from the energetic burden of unnecessary protein production, which was found to delay growth [10–12, 48].

It was previously shown that transcriptional error for a single gene is minimized when its binding site is occupied [7]—a regulatory strategy equivalent to the Savageau demand rule. However, single-gene models neglected the increase in erroneous interactions that can occur

following network augmentation beyond the single gene. The regulatory cost increases super-linearly with the number of molecular species and regulatory interactions and can therefore only be determined when the network is considered as whole. This would result in a different mathematical solution to minimize global crosstalk, compared to the single-gene case. For comparison between single-gene and global crosstalk models see S1 Text (Section 9).

Selection to reduce global regulatory crosstalk [39, 49], was reported in previous bioinformatic studies. Our finding that effective similarity obtained for the *S. cerevisiae* gene regulatory network is lower than the median effective similarity obtained in random networks with similar parameters, corroborated these reports (Fig 4C). Yet, crosstalk is not fully eliminated by selection. Despite the functional interference it causes in the short run, crosstalk is thought to promote evolvability in both gene regulatory and signaling networks in the long run [50–54]. However, the interplay between these two opposing effects of crosstalk, is still poorly understood.

Crosstalk reduction is one of several functional considerations shaping the evolution of gene regulatory networks. Other considerations include the network dynamical properties [55] and protein production requirements [6]. Above all, evolution is a random process and certain network designs become fixed and continue propagating [56–59]. For example, new transcription factors often evolve by duplication of an existing TF followed by sub- or neo-functionalization, thereby preserving the form of regulation of the ancestral TF [60]. Taken together, a generalized model for network evolution, which would incorporate the effects of crosstalk on different time scales, alongside traditional selection on the network to achieve a certain input-output goal, remains to be formulated.

Methods

Distribution of t is approximated by a Gaussian distribution

Given that the cognate TF of gene i is 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 with probabilities γ_i, that are not necessarily identically distributed. Its mean and variance are:

$$\langle t \rangle = \frac{1}{M}\sum_{i=1}^{M}\gamma_i = \langle \gamma_i \rangle, \tag{8}$$

$$\mathrm{var}(t) = \frac{1}{M}\sum_{i=1}^{M}\gamma_i(1-\gamma_i) = \langle \gamma_i(1-\gamma_i) \rangle. \tag{9}$$

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

Exact solution of the 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:

$$f_{X^*}(X^*) = \sum_i f_t(g_i^{-1}(X^*))\left|\frac{\mathrm{d}g_i^{-1}(X^*)}{\mathrm{d}X^*}\right|, \tag{10}$$

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

branches:

$$f_{X^*}(X^*) = f_t(g_1^{-1}(X^*))\left|\frac{\mathrm{d}g_1^{-1}(X^*)}{\mathrm{d}X^*}\right| + f_t(g_2^{-1}(X^*))\left|\frac{\mathrm{d}g_2^{-1}(X^*)}{\mathrm{d}X^*}\right|. \quad (11)$$

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

For regime I, the lower limit on crosstalk is $X^*(t) = t$. Its inverse is $g^{-1}(X^*) = t(X^*) = X^*$, while the derivative $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = 1$. Similarly, in regime II, the lower limit on 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 derivative $\mathrm{d}g^{-1}(X^*)/\mathrm{d}X^* = -(1 - \alpha - \alpha X^*)^{-2}$. The analytical solution for regime III was computed using Mathematica and the solution can be found in S2 Appendix.

Using these values, one can compute $f_{X^*}(X^*)$ for X^* in all three regimes.

Stochastic semi-analytical solution of crosstalk for a random number of present TFs

For each gene i, we randomly draw, 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 in the relevant regime (I, II or III). 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 summary statistics of TF binding site sequences. Each element c_{ij} designates the number of known TF binding site sequences with nucleotide i in position j. We obtained the PCMs from the 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 computed the energy mismatch value as $\epsilon_{ij} = \log(\frac{c_{mj}}{c_{ij}})$, where $c_{mj} = \max_i c_{ij}$ is the maximal count at position j. To avoid divergence of the energy ϵ_{ij} in case of zero counts, $c_{ij} = 0$, we added a constant pseudocount $\delta = 0.1$ to all matrix entries.

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 was 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 applied a commonly used practice of adding a pseudocount δ to all PCM entries. Following a previous work [22], where various δ values were compared to an information method (where pseudocount is not needed), we set $\delta = 0.1$.

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

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

Numerical computation of similarity measure using PCMs

To compute the similarity measure between binding site k and a transcription factor l, we first substituted 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 distinguished between the following cases:

- If the TF consensus sequence l was shorter than the binding site sequence k, we computed the energies for all possible overlaps of the shorter sequence with respect to the longer one. We took the minimal value to be the binding energy.
- If the TF consensus sequence l was longer than the binding site sequence k, the TF energy matrix was again slid along the binding site and energies were calculated again for every relative positioning of the two sequences. The only difference from the previous case was that energetic contributions from positions where the TF binds outside the binding site, were taken into account by averaging energies over all four nucleotides. The 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 the 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. Here too, we computed the binding energy for all possible overlaps between the BS and TF and took the lowest value 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 $E^{kl} \neq E^{lk}$. The similarity measure between binding site k and all other binding sites was computed as the average Boltzmann weight, taken over all non-cognate TF binding to binding site k:

$$S_k = \frac{1}{T} \sum_{l=1,\ l \neq k}^{M} C_l e^{-E^{kl}}, \tag{12}$$

with C_l being the concentration of TF species l, and T the number of present TF species.

Numerical computation of crosstalk given PCMs

For the numerical computation of crosstalk, we used the matrix of binding energies E^{kl} between binding site k and TF l, using the following algorithm:

1. randomly choose a subset of genes that should be regulated by their cognate TF. At each realization, a different subset is chosen. All subsets form a proportion t of the genes.
2. For gene k, obtain the similarity measure $S_k = \frac{1}{T} \sum_{l \neq k} C_l e^{-E^{kl}}$. Set the concentration of the absent TFs to zero, and set equal concentrations ($C_l = C$) to all present TFs, 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. Obtain the total crosstalk X by summing over the contributions of all individual genes.
5. Average over a large number of realizations (we used several hundred realizations for which the average crosstalk had 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 a gene could be regulated by multiple TFs

In an actual gene regulatory network, many TFs regulate multiple genes and many genes are regulated by multiple TFs rather than the one-to-one TF-gene association we considered so far. Specifically, in our data, around 96% of the TFs regulate more than one gene. To account for that, we obtained the list of genes that are regulated by the given *S. cerevisiae* transcription factors [38]. Numerical crosstalk calculation for this network closely followed the previous procedure. The only difference was the computation of the similarity measure of genes regulated by multiple cognate TFs. Such genes have multiple binding site sequences (one for each cognate TF) and consequently, multiple binding energies and similarity measures. We then calculated a unified similarity measure per gene as follows:

1. For a given gene k, find all the TFs that regulate it.
2. Obtain the consensus sequences of these TFs.
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. 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 12.
5. Use the mean of the computed S_{k^i} similarity measures taken over the various binding sites of gene k as the unified similarity of that gene.

Simulating synthetic data

To simulate synthetic data of TF binding preferences, we constructed artificial PCMs, using the data of the 23 yeast energy matrices, as follows. We first created the nucleotide abundance distribution of the yeast TFs consensus sequence and then drew random realizations from this distribution to obtain a consensus sequence for each synthetic TF. This distribution was non-uniform and biased towards excess of A and T nucleotides. We allowed for a variety of consensus sequence lengths, using the same length distribution as in the yeast data. Similarly, we created the distribution of the non-consensus energy values of the 23 TFs energy matrices and drew random realizations from this distribution to construct the energy matrices for the synthetic TFs.

Computing the subnetworks of synthetic data and their crosstalk

To construct a full network, we fabricated data for 300 TFs, as described above. We then computed the network's matrix of binding energies $E^{kl}_{\text{full network}}$ of the l-th TF to the k-th binding site,

where the each binding site sequence was taken as the consensus sequence of its cognate TF, as in the yeast data. We next formed subnetworks of this full network, by choosing a subset of TFs and taking the corresponding subset of binding energy entries, to obtain $E^{kl}_{\text{subnetwork}}$. We used either randomly chosen subsets of TFs ("random networks") or deterministically picked the subset of TFs having the highest similarity measure S_i^{full} network with respect to the full network. We then numerically computed minimal crosstalk X^* for each subnetwork, following the same procedure as for the yeast data. We repeated this procedure for 100 randomly drawn full networks.

Comparison of the numerical results to the analytical expression

We fit the analytical expression for $X^*(t)$ to the numerically calculated crosstalk. The main difference between the two approaches is that the analytical expression assumes uniform S_k values for all TFs, whereas the numerical approach allows for diverse S_k values. We assumed a single representative $s_{\text{effective}}$ value that would best fit the numerical result. For this, we minimized the sum of squared differences over various values of t to find the best $s_{\text{effective}}$. Distributions of $s_{\text{effective}}$ values were based on 100 randomly drawn full networks from which subnetworks were sampled. For each subnetwork size, we sampled each of the full networks just once, to avoid correlations between the random subnetworks.

Supporting information

S1 Text.
(PDF)

S1 Appendix. Maximizing crosstalk. Mathematica notebook, showing which argument maximizes crosstalk.
(NB)

S2 Appendix. Distribution of crosstalk values. Mathematica notebook, showing computation of distribution of crosstalk values.
(NB)

Acknowledgments

We thank Uri Alon and Tzachi 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. R. G. thanks Kirti Jain for discussions.

Author Contributions

Conceptualization: Tamar Friedlander.

Formal analysis: Rok Grah.

Investigation: Rok Grah.

Supervision: Tamar Friedlander.

Writing – original draft: Rok Grah, Tamar Friedlander.

Writing – review & editing: Rok Grah, Tamar Friedlander.

References

1. Savageau MA. Genetic Regulatory Mechanisms and the Ecological Niche of Escherichia coli. Proceedings of the National Academy of Sciences. 1974 Jun; 71(6):2453–2455. https://doi.org/10.1073/pnas.71.6.2453
2. Savageau MA. Design of molecular control mechanisms and the demand for gene expression. Proceedings of the National Academy of Sciences. 1977 Dec; 74(12):5647–5651. https://doi.org/10.1073/pnas.74.12.5647
3. Savageau MA. Regulation of differentiated cell-specific functions. Proceedings of the National Academy of Sciences. 1983 Mar; 80(5):1411–1415. https://doi.org/10.1073/pnas.80.5.1411
4. Gerland U, Hwa T. Evolutionary selection between alternative modes of gene regulation. Proceedings of the National Academy of Sciences. 2009 Jun; 106(22):8841–8846. https://doi.org/10.1073/pnas.0808500106
5. Savageau MA. Demand Theory of Gene Regulation. I. Quantitative Development of the Theory. Genetics. 1998 Aug; 149(4):1665–1676. PMID: 9691027
6. Kumar Prajapat M, Jain K, Choudhury D, Raj N, Saini S. Revisiting demand rules for gene regulation. Molecular BioSystems. 2016; 12(2):421–430. https://doi.org/10.1039/C5MB00693G
7. Shinar G, Dekel E, Tlusty T, Alon U. Rules for biological regulation based on error minimization. Proceedings of the National Academy of Sciences of the United States of America. 2006 Mar; 103(11):3999–4004. https://doi.org/10.1073/pnas.0506610103 PMID: 16537475
8. Sasson V, Shachrai I, Bren A, Dekel E, Alon U. Mode of Regulation and the Insulation of Bacterial Gene Expression. Molecular Cell. 2012 May; 46(4):399–407. https://doi.org/10.1016/j.molcel.2012.04.032 PMID: 22633488
9. Novick A, Weiner M. Enzyme Induction as an All-or-None Phenomenon. Proceedings of the National Academy of Sciences. 1957 Jul; 43(7):553–566. https://doi.org/10.1073/pnas.43.7.553
10. Koch AL. The protein burden oflac operon products. Journal of Molecular Evolution. 1983 Nov; 19(6):455–462. https://doi.org/10.1007/bf02102321 PMID: 6361271
11. Kurland CG, Dong H. Bacterial growth inhibition by overproduction of protein. Molecular Microbiology. 1996 Jul; 21(1):1–4. https://doi.org/10.1046/j.1365-2958.1996.5901313.x PMID: 8843428
12. Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005 Jul; 436(7050):588–592. https://doi.org/10.1038/nature03842 PMID: 16049495
13. Kafri M, Metzl-Raz E, Jona G, Barkai N. The Cost of Protein Production. Cell Reports. 2016 Jan; 14(1):22–31. https://doi.org/10.1016/j.celrep.2015.12.015 PMID: 26725116
14. Von Hippel PH, Revzin A, Gross CA, Wang AC. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects. Proceedings of the National Academy of Sciences. 1974; 71(12):4808–4812. https://doi.org/10.1073/pnas.71.12.4808
15. Johnson JM, Edwards S, Shoemaker D, Schadt EE. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends in Genetics. 2005 Feb; 21(2):93–102. https://doi.org/10.1016/j.tig.2004.12.009 PMID: 15661355
16. Maerkl SJ, Quake SR. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science. 2007 Jan; 315(5809):233–237. https://doi.org/10.1126/science.1131007 PMID: 17218526
17. Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends in Genetics. 2009 Oct; 25(10):434–440. https://doi.org/10.1016/j.tig.2009.08.003 PMID: 19815308
18. Rockel S, Geertz M, Hens K, Deplancke B, Maerkl SJ. iSLIM: a comprehensive approach to mapping and characterizing gene regulatory networks. Nucleic acids research. 2012;p. gks1323. https://doi.org/10.1093/nar/gks1323 PMID: 23258699
19. Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. Nature Communications. 2018 Apr; 9(1):1530. https://doi.org/10.1038/s41467-018-04026-w PMID: 29670097
20. Gerland U, Moroz JD, Hwa T. Physical constraints and functional characteristics of transcription factor–DNA interaction. Proceedings of the National Academy of Sciences. 2002; 99(19):12015–12020. https://doi.org/10.1073/pnas.192693599
21. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: models. Current Opinion in Genetics & Development. 2005 Apr; 15(2):116–124. https://doi.org/10.1016/j.gde.2005.02.007
22. Friedlander T, Prizak R, Guet CC, Barton NH, Tkačik G. Intrinsic limits to gene regulation by global crosstalk. Nature Communications. 2016 Aug; 7:12307. https://doi.org/10.1038/ncomms12307 PMID: 27489144

23. Shea MA, Ackers GK. The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation. Journal of Molecular Biology. 1985 Jan; 181(2):211–230. https://doi.org/10.1016/0022-2836(85)90086-5 PMID: 3157005

24. Von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. Proceedings of the National Academy of Sciences. 1986; 83(6):1608. https://doi.org/10.1073/pnas.83.6.1608

25. Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences. 2010 May; 107(20):9158–9163. https://doi.org/10.1073/pnas.1004290107

26. Lässig M. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics. 2007; 8(6):1–21.

27. Haldane A, Manhart M, Morozov AV. Biophysical Fitness Landscapes for Transcription Factor Binding Sites. PLOS Computational Biology. 2014 Jul; 10(7):e1003683. https://doi.org/10.1371/journal.pcbi.1003683 PMID: 25010228

28. Sarai A, Takeda Y. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proceedings of the National Academy of Sciences. 1989 Sep; 86(17):6513–6517. https://doi.org/10.1073/pnas.86.17.6513

29. Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, et al. *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nature Biotechnology. 2010 Sep; 28(9):970–975. https://doi.org/10.1038/nbt.1675 PMID: 20802496

30. Afek A, Schipper JL, Horton J, Gordân R, Lukatsky DB. Protein-DNA binding in the absence of specific base-pair recognition. Proceedings of the National Academy of Sciences. 2014 Oct;p. 201410569.

31. Landman J, Brewster RC, Weinert FM, Phillips R, Kegel WK. Self-consistent theory of transcriptional control in complex regulatory architectures. PLOS ONE. 2017 Jul; 12(7):e0179235. https://doi.org/10.1371/journal.pone.0179235 PMID: 28686609

32. Itzkovitz S, Tlusty T, Alon U. Coding limits on the number of transcription factors. BMC Genomics. 2006 Sep; 7(1):239. https://doi.org/10.1186/1471-2164-7-239 PMID: 16984633

33. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell. 2005 Dec; 123(6):1025–1036. https://doi.org/10.1016/j.cell.2005.09.031 PMID: 16360033

34. Wang Y, Guo L, Golding I, Cox EC, Ong NP. Quantitative Transcription Factor Binding Kinetics at the Single-Molecule Level. Biophysical Journal. 2009 Jan; 96(2):609–620. https://doi.org/10.1016/j.bpj.2008.09.040 PMID: 19167308

35. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Molecular Biology of the Cell. 2000 Dec; 11(12):4241–4257. https://doi.org/10.1091/mbc.11.12.4241 PMID: 11102521

36. Spivak AT, Stormo GD. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Research. 2012 Jan; 40(D1):D162–D168. https://doi.org/10.1093/nar/gkr1180 PMID: 22140105

37. Ghaemmaghami S, Huh WK, Bower K, Howson RW, others. Global analysis of protein expression in yeast. Nature. 2003; 425(6959):737. https://doi.org/10.1038/nature02046 PMID: 14562106

38. Saccharomyces Genome Database | SGD;.

39. Qian L, Kussell E. Genome-Wide Motif Statistics are Shaped by DNA Binding Proteins over Evolutionary Time Scales. Physical Review X. 2016 Oct; 6(4):041009. https://doi.org/10.1103/PhysRevX.6.041009

40. Carter MG, Sharov AA, VanBuren V, Dudekula DB, Carmack CE, Nelson C, et al. Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. Genome Biology. 2005 Jun; 6:R61. https://doi.org/10.1186/gb-2005-6-7-r61 PMID: 15998450

41. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014 Feb; 11(2):163–166. https://doi.org/10.1038/nmeth.2772 PMID: 24363023

42. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Molecular Systems Biology. 2011 Jan; 7(1):548. https://doi.org/10.1038/msb.2011.81 PMID: 22068331

43. Tao H, Bausch C, Richmond C, Blattner FR, Conway T. Functional Genomics: Expression Analysis ofEscherichia coli Growing on Minimal and Rich Media. Journal of Bacteriology. 1999 Oct; 181(20):6425–6440. https://doi.org/10.1128/JB.181.20.6425-6440.1999 PMID: 10515934

44. Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA, LaRossa RA. High-Density Microarray-Mediated Gene Expression Profiling of Escherichia coli. Journal of Bacteriology. 2001 Jan; 183(2):545–556. https://doi.org/10.1128/JB.183.2.545-556.2001 PMID: 11133948

45. Genes IX 9th edition by Lewin, Benjamin published by Jones & Bartlett Publishers Hardcover. Jones & Bartlett Publishers; 2007.

46. Price MN, Deutschbauer AM, Skerker JM, Wetmore KM, Ruths T, Mar JS, et al. Indirect and suboptimal control of gene expression is widespread in bacteria. Molecular Systems Biology. 2013 Jan; 9(1):660. https://doi.org/10.1038/msb.2013.16 PMID: 23591776

47. Hershberg R, Margalit H. Co-evolution of transcription factors and their targets depends on mode of regulation. Genome Biology. 2006; 7:R62. https://doi.org/10.1186/gb-2006-7-7-r62 PMID: 16859509

48. Shachrai I, Zaslaver A, Alon U, Dekel E. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. Molecular Cell. 2010 Jun; 38(5):758–767. https://doi.org/10.1016/j.molcel.2010.04.015 PMID: 20434381

49. Hahn MW, Stajich JE, Wray GA. The Effects of Selection Against Spurious Transcription Factor Binding Sites. Molecular Biology and Evolution. 2003 Jun; 20(6):901–906. https://doi.org/10.1093/molbev/msg096 PMID: 12716998

50. Shultzaberger RK, Maerkl SJ, Kirsch JF, Eisen MB. Probing the Informational and regulatory plasticity of a transcription factor DNA–binding domain. PLoS genetics. 2012; 8(3). https://doi.org/10.1371/journal.pgen.1002614 PMID: 22496663

51. Payne JL, Wagner A. The Robustness and Evolvability of Transcription Factor Binding Sites. Science. 2014 Feb; 343(6173):875–877. https://doi.org/10.1126/science.1249046 PMID: 24558158

52. Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. Cell. 2015 Oct; 163(3):594–606. https://doi.org/10.1016/j.cell.2015.09.055 PMID: 26478181

53. Friedlander T, Prizak R, Barton NH, Tkačik G. Evolution of new regulatory functions on biophysically realistic fitness landscapes. Nature Communications. 2017 Aug; 8(1):216. https://doi.org/10.1038/s41467-017-00238-8 PMID: 28790313

54. Rowland MA, Greenbaum JM, Deeds EJ. Crosstalk and the evolvability of intracellular communication. Nature Communications. 2017 Jul; 8:ncomms16009. https://doi.org/10.1038/ncomms16009

55. Alon U. Network motifs: theory and experimental approaches. Nature Reviews Genetics. 2007 Jun; 8(6):450–461. https://doi.org/10.1038/nrg2102 PMID: 17510665

56. Wagner A. Robustness and evolvability: a paradox resolved. Proceedings of the Royal Society B: Biological Sciences. 2008; 275(1630):91–100. https://doi.org/10.1098/rspb.2007.1137 PMID: 17971325

57. Fontana W, Buss LW. What would be conserved if "the tape were played twice"? Proceedings of the National Academy of Sciences. 1994 Jan; 91(2):757–761. https://doi.org/10.1073/pnas.91.2.757

58. Friedlander T, Mayo AE, Tlusty T, Alon U. Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems. PLoS ONE. 2013 Aug; 8(8). https://doi.org/10.1371/journal.pone.0070444

59. Martin OC, Krzywicki A, Zagorski M. Drivers of structural features in gene regulatory networks: From biophysical constraints to biological function. Physics of Life Reviews. 2016 Jul; 17:124–158. https://doi.org/10.1016/j.plrev.2016.06.002 PMID: 27365153

60. Nguyen CC, Saier MH. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. FEBS Letters. 1995 Dec; 377(2):98–102. https://doi.org/10.1016/0014-5793(95)01344-x PMID: 8543068