How gene order is influenced by the biophysics of transcription regulation

Grigory Kolesov*, Zeba Wunderlich†, Olga N. Laikova‡, Mikhail S. Gelfand§, and Leonid A. Mirny*¶

*Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139; †Biophysics Program, Harvard University, Cambridge, MA 02138; ‡State Scientific Center GosNIIGenetika, Moscow 117545, Russia; and §Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved July 2, 2007 (received for review January 24, 2007)

What are the forces that shape the structure of prokaryotic genomes: the order of genes, their proximity, and their orientation? Coregulation and coordinated horizontal gene transfer are believed to promote the proximity of functionally related genes and the formation of operons. However, forces that influence the structure of the genome beyond the level of a single operon remain unknown. Here, we show that the biophysical mechanism by which regulatory proteins search for their sites on DNA can impose constraints on genome structure. Using simulations, we demonstrate that rapid and reliable gene regulation requires that the transcription factor (TF) gene be close to the site on DNA the TF has to bind, thus promoting the colocalization of TF genes and their targets on the genome. We use parameters that have been measured in recent experiments to estimate the relevant length and times scales of this process and demonstrate that the search for a cognate site may be prohibitively slow if a TF has a low copy number and is not colocalized. We also analyze TFs and their sites in a number of bacterial genomes, confirm that they are colocalized significantly more often than expected, and show that this observation cannot be attributed to the pressure for coregulation or formation of selfish gene clusters, thus supporting the role of the biophysical constraint in shaping the structure of prokaryotic genomes. Our results demonstrate how spatial organization can influence timing and noise in gene expression.

diffusion | genetics | genomics | protein–DNA interactions | spatial effects

The colocalization of prokaryotic transcription factor (TF) genes and their binding sites is known from the pioneering work of Jacob and Monod (1) on the lactose operon and has been shown to be widespread (2–4) and essential for the formation of regulatory motifs (5). Some have hypothesized that TF-binding site colocalization is advantageous, in part, because it could expedite a TF's search for its site (2, 5–7) (the rapid search hypothesis). In prokaryotes, this speed-up by colocalization is possible because transcription and translation are coupled spatially and temporally. Therefore, TFs are synthesized near their genes and can rapidly bind colocalized sites (Fig. 1*A*). The arrival time of a TF to its site ultimately controls the timing of gene regulation, whereas fluctuations in the arrival time can lead to bursts of gene activity and noise in gene regulation. The rapid search hypothesis suggests that colocalization is favorable because expediting TF arrival makes regulation faster and more reliable.

Both experimentally (see ref. 8 for an overview) and theoretically (9–13), many have studied the broader question: how can a TF find its cognate site on DNA among $\approx 10^7$ decoy sites in a fraction of a minute while moving in the crowded environment of the cell and hampered by other DNA-bound proteins? The general model of the process includes 3D spatial diffusion of the TF through the cell volume and 1D sliding of a TF along DNA. According to this model, the search process consists of multiple rounds of search, alternating between 1D sliding and 3D spatial diffusion, leading to the expression for the mean search time, t_s, obtained (in different forms) by several groups (9–13):

$$t_s = \frac{M}{s}(\tau_{1D} + \tau_{3D}), \qquad [1]$$

where M is the total length of DNA in the cell, s is the sliding length, i.e., the mean number of base pairs scanned in a single round of sliding, and τ_{1D} and τ_{3D} are the mean durations of a single round of 1D sliding and 3D diffusion, respectively. However, it is not intuitively clear why colocalization would cause a speed-up, because in Eq. **1**, as in traditional reaction rate theory, the search (reaction) time is distance-independent. The distance (and time) independence of the reaction rate is characteristic of 3D systems, whereas reactions in 2D and 1D systems are distance-dependent (14).

Here, we systematically investigate the rapid search hypothesis and assess it against the alternative but complementary views that colocalization is due to coregulation or self-regulation or to enable horizontal transfer of functionally coupled genes (the selfish gene cluster hypothesis) (15, 16). We approach the problem by taking the following three steps: we (*i*) estimate the TF search time in bacteria and determine the degree of acceleration provided by TF-binding site colocalization, (*ii*) estimate the extent of colocalization in bacterial genomes, and (*iii*) consider and rule out alternative explanations of colocalization. We demonstrate that the requirement for rapid search imposes a significant constraint on the evolution of gene order, an interesting case where a biophysical mechanism influences genome organization.

Results

How Much Acceleration Can Be Achieved by Colocalization? To connect the search time calculations to DNA conformation, we note that Eq. **1** implicitly assumes that each round of sliding is independent: the rounds of 3D diffusion between the slide completely randomize the position of the TF. To relax this assumption, we considered two types of 3D motion: small hops and large-scale jumps (Fig. 1*B*). Hops are rapid reassociations of a TF to the same region of DNA. Elegant biochemical experiments have demonstrated hopping of DNA-binding proteins on DNA (17). We found that hops results from the geometry of the problem: Once a TF dissociates from DNA, it is much more likely to associate again to the same region of DNA than to other remote strands. We also dem-

Author contributions: G.K. and Z.W. contributed equally to this work; G.K., Z.W., and L.A.M. designed research; G.K., Z.W., and L.A.M. performed research; O.N.L. and M.S.G. contributed new reagents/analytic tools; G.K., Z.W., O.N.L., M.S.G., and L.A.M. analyzed data; and G.K., Z.W., and L.A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: BS, binding site; TF, transcription factor; TU, transcription unit.

¶To whom correspondence should be addressed at: 77 Massachusetts Avenue, 16-343, Cambridge, MA 02139. E-mail: leonid@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0700672104/DC1.

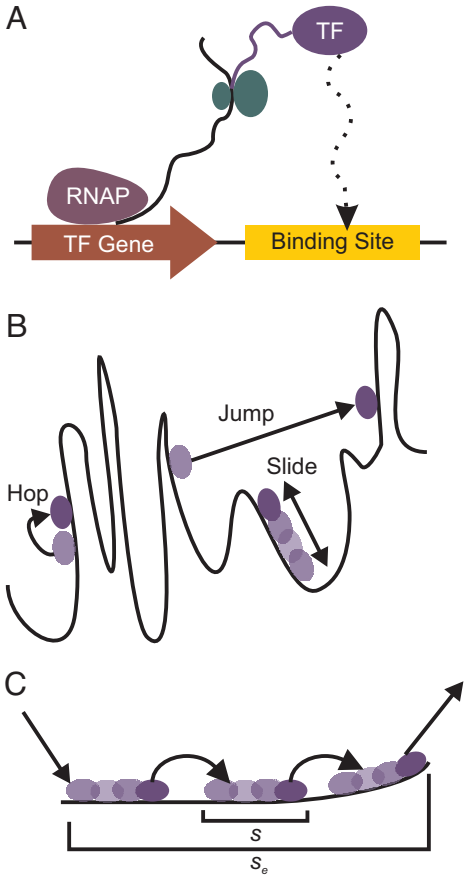

Fig. 1. We propose the rapid search hypothesis as an explanation for colocalization of transcription factor genes and their targets and model the search process with hops, jumps, and slides. (*A*) The rapid search hypothesis. In prokaryotes, transcription and translation are coupled; therefore, transcription factors (TFs) are released from the ribosome near their encoding gene, enabling a TF to rapidly search the DNA nearby. The rapid search hypothesis suggests that TF genes and their binding sites may be colocalized on the chromosome because this enables newly synthesized TFs to rapidly find their binding sites. (*B*) Model of the transcription factor search process. We define three types of movements for TFs: slides, rounds of 1D diffusion along the DNA; hops, short rounds of 3D diffusion where the TF dissociates from the DNA and rebinds at a site very nearby; and jumps, longer rounds of 3D diffusion where the TF dissociates from the DNA and binds a site that may be quite far away. Mathematically, the dissociation and association sites of hops are correlated, whereas those of jumps are uncorrelated. We then model the search process as alternating rounds of 3D and 1D diffusion; the TF ends the slide with either a hop or a jump. (*C*) We find the hops are so short that they can be accounted for by rescaling the sliding length, *s*, the number of base pairs scanned in a slide by the number of hops per jump, n_{hops}, to get s_e, the number of base pairs scanned in between jumps: $s_e = s\cdot\sqrt{n_{hops}}$.

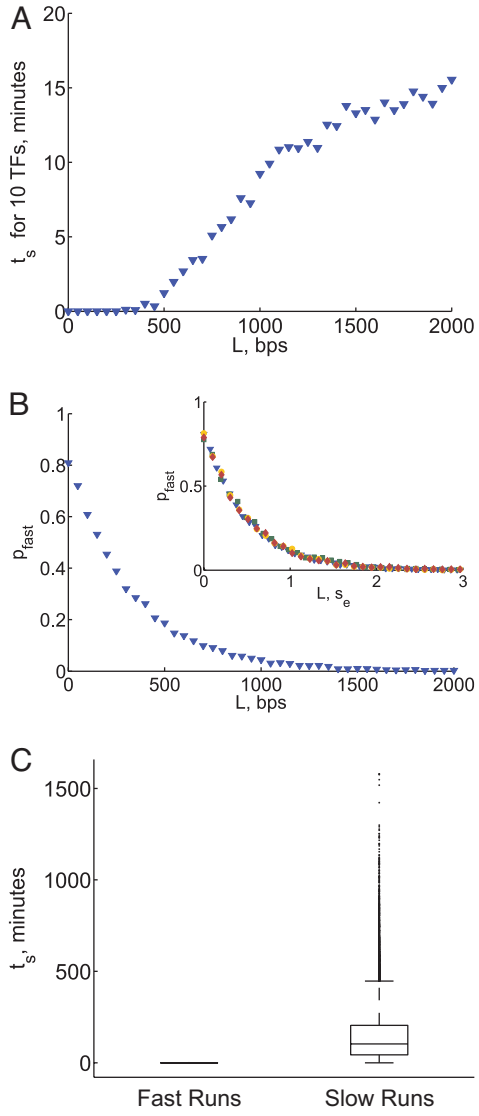

Fig. 2. Simulations of the transcription factor search process show that its length depends on starting point. (*A*) Search time for a group of 10 TFs versus *L*. Here, we simulated a group of 10 TFs searching for a binding site and plot the mean search time, t_s, of the first TF to reach the site versus initial distance *L*. Here, s_e = 660 bp, and 500 runs were simulated for each *L*. (*B*) The probability of fast runs. Here, the probability of a fast run, a run in which the TF starts near its binding site and finds it by hopping and sliding but without jumping, is plotted versus the initial distances between the TF and its site, *L*. The main plot shows *L* in base pairs, and the value of $s_e = s\cdot\sqrt{n_{hops}} \approx 660$ bp. (*Inset*) *L* in units of s_e and the different symbols correspond to different values of *s* (blue triangles, *s* = 270; green squares, *s* = 50; yellow circles, *s* = 100; red diamonds, *s* = 500). Each data point is the mean of 1,000 trials. The overlap in *Inset* shows that the behavior is parameter-independent when *L* is expressed in units of s_e, confirming that s_e is the only relevant parameter in this simulation. (*C*) Distribution of run times for fast and slow runs. The distribution of search times, t_s is plotted for fast and slow runs, where fast runs are defined as above and slow runs are searched where the TF uses hopping, sliding, and jumping to find its binding site. The box has lines at the lower, median, and upper quartile values. The whiskers extend from the box to 1.5 times the interquartile range, the difference between the lower and upper quartiles. Data points beyond the whiskers are noted as circles. Each plot includes the data from 30,000 runs. This plot clearly shows that (*i*) fast runs are much faster than slow runs (fast runs have a median of 0.2 s, whereas slow runs have a median of 100 min) and (*ii*) fast runs are much less variable than slow runs.

onstrated that hops are short and can be accounted for by replacing the sliding length *s* by an effective sliding length $s_e = s\cdot\sqrt{n_{hops}}$, where n_{hops} is the mean number of hops a TF makes before a jump (Fig. 1*C*). Using simulations of spatial diffusion through a realistic geometry and density of nonspecific DNA, we estimated $n_{hops} \approx 5$–6 (46).

Using simulations, we calculated search time as a function of the initial distance between a TF and its site (*L*). Here, we observe two types of searches. When released from the ribosome, a TF can bind DNA near the 3′ end of its gene and start sliding and hopping along DNA. If the cognate site is reached this way, the average search time is fast (≈0.3 sec; Fig. 2*C*). Alternatively, if a TF dissociates from DNA and jumps before binding its site, then it must sample the whole genome to find its site, and the search is slow (≈150 min; Fig. 2*C*). The choice between these scenarios is controlled by a single length scale, the effective sliding length $s_e = s\cdot\sqrt{n_{hops}} \approx 660$ bp, with a range

between 70 and 2,000 bp, for a typical TF. Sites at distance $L < s_e$ are likely to be found quickly, whereas more distant site require a slow global search. Fig. 2*A* shows the average search time (t_s) for 10 TFs, Fig. 2*B* shows the probability of a fast search (p_{fast}) as a function of L, and Fig. 2*C* shows the distribution of times of fast and slow runs (46).

Connecting back to the theory, our slow searches are described by Eq. **1**. But why are they so slow? Although the form of Eq. **1** is intuitive, it does not show how the value of t_s depends on the physical properties of the system. The sliding length s determines the number of rounds of search needed to find the slide. The search time also depends on the ratio of the time spent on the DNA to the time spent in the cytoplasm: τ_{1D}/τ_{3D}. This ratio is controlled by the affinity of a TF for nonspecific DNA, K_d^{NS}, and the total concentration of nonspecific DNA in the cell, [DNA]. Although sliding can increase the rate of search by reducing the number of rounds of search, it requires a TF to have an affinity for nonspecific DNA, which in turn can slow down search. The balance between these factors controls the global efficiency of search. To show these dependencies, Eq. **1** can be written in the following form [see supporting information (SI) *Text*]:

$$t_s = \frac{M}{k_{on}[\mathrm{DNA}]} \cdot \frac{\left(1 + \frac{[\mathrm{DNA}]}{K_d^{NS}}\right)}{s}, \qquad [2]$$

where the first term is the search time in the absence of sliding and nonspecific binding, whereas the second term provides the balance between the speed-up due to sliding ($1/s$) and the slow-down due to the nonspecific binding (the ratio of K_d^{NS} and [DNA]). Note that 3D and 1D diffusion coefficients are agglomerated into k_{on}, the on rate of a TF to bind DNA by a spatial diffusion [the Smoluchowski rate (18)], and s, respectively. As we showed earlier (10), search time is minimized when equal time is spent on DNA and in the solvent (i.e., $\tau_{1D} = \tau_{3D}$). However, *in vivo*, the strong affinity for nonspecific DNA [$K_d^{NS} \approx 10^{-3}$ to 10^{-6} M (19)] and the high concentration of DNA inside the cell [[DNA] = 10^{-2} M (20)] cause TFs to spend a significant amount of their time on nonspecific DNA ($\tau_{1D}/\tau_{3D} = [\mathrm{DNA}]/K_d^{NS} \approx 10^1$ to 10^4). This nonoptimal time partitioning leads to search times from 15 to 500 min for a single TF.

Clearly, having multiple copies of a TF significantly speeds up the search (linear with the number of copies). However, available *in vivo* measurements suggest there are only ≈10 copies of lactose repressor per cell (21), whereas there are >200 copies of ArcA per cell (22), a global regulator with >50 targets in the cell.

Therefore, the acceleration of binding provided by colocalization can have a significant effect on gene regulation for low-copy-number TFs. If the TF is a repressor, rapid binding leaves little time for a polymerase to bind a promoter and start transcription, so bursts of gene activity are short and rare, consistent with recent single-molecule experiments (23, 24). However, if it takes ≈15 min for a pool of ≈10 repressors to bind a site (Fig. 2*A*), the bursts of gene activity are long, making repression leaky and inefficient. Slow searches make the time required for transcription regulation comparable with the duplication time of bacteria, thus putting slowly regulating bacteria at significant disadvantage.

To summarize, simulations show that TF binding is slow if TFs are not colocalized and have low copy number. Rapid search can be achieved by either colocalization or by increasing the copy number of each TF, arguably a more costly solution. Therefore, colocalization provides a significant advantage for low-copy-number TFs.

How Widespread Is Colocalization That Cannot Be Attributed to Co/Self-Regulation in Bacteria? To unravel the extent of colocalization, we examined the distances between LacI/GalS family TFs and their binding sites. We grouped TFs into two categories: global TFs (25, 26), which are pleiotropic and regulate more than four operons (FruR, PurR, and CcpA), and local TFs, which regulate fewer than four operons. To focus on colocalization because of rapid search, we excluded from consideration all sites that can have a role in coregulation of the TF and its regulated transcription units (TUs) or self-regulation of the TF (Fig. 3*A*).

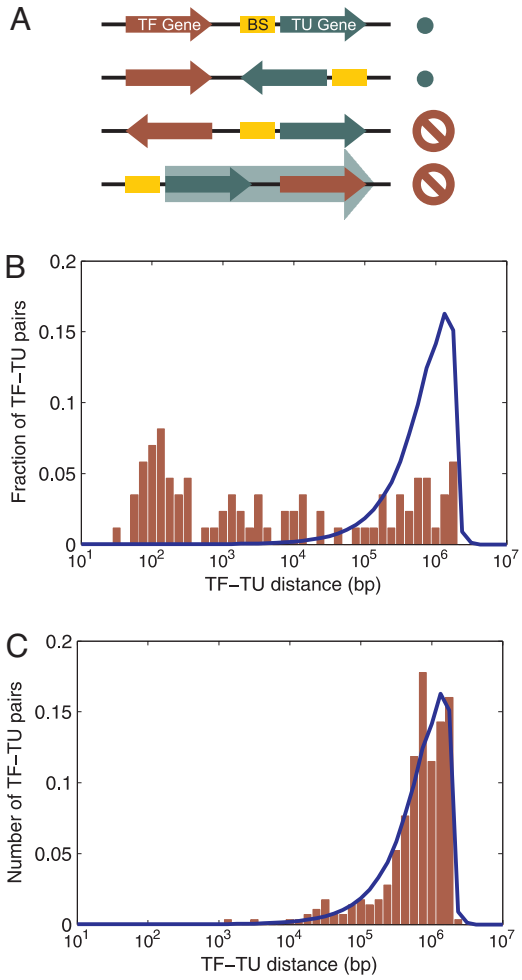

Fig. 3. We show that local transcription factors are colocalized with their targets. (*A*) Possible orientations of a TF gene, its BS, and the regulated TU. In this diagram, the TF gene encodes a TF that regulates the expression of the TU by binding the BS. In our study, we aimed to determine the extent of colocalization that cannot be explained by co- or self-regulation. Therefore, we excluded the third orientation, because the TF and TU may be coregulated through a shared promoter region (coregulation), and the fourth orientation, because the TF may be part of the same operon as the TU (self-regulation). (*B*) Distances between local TFs and their binding sites. Here, TF–TU distances for local regulators are shown as bars, and the distances expected from random TF–TU assignments are shown by the blue line. Here, we can see that local TFs are significantly colocalized with their binding sites on length scales comparable with $s_e \approx 10^3$ bps, suggesting that the rapid search hypothesis is feasible. (*C*) Distances between global TFs and their binding sites. Here again, TF–TU distances for global regulators are shown as bars, and the expected distribution for the random TF–TU assignment is shown by the blue line. For global TFs, there is no significant colocalization, suggesting that rapid search may be achieved by high copy number instead.

Fig. 3 presents the distribution of the distance between TFs and their TUs for local and global TFs. Each distribution is compared with expected distribution of distances between random locations on chromosomes. The distribution for local TFs (Fig. 3*B*) is strikingly different from those of global TFs (Fig. 3*C*)

and the expected random distribution; there is a significant excess of local TF-site pairs at distances below $s_e \approx 1{,}000$ bp. This result clearly demonstrates that widespread colocalization of TFs and their sites cannot be explained solely by requirements for co- or self-regulation (2, 5, 27). Global regulators exhibit no significant colocalization, suggesting that they must achieve a speed-up in recognition by being present in high copy numbers, a claim supported by experimental evidences. For example, under normal growth conditions, the estimated transcript number of the global regulator *crp* is two orders of magnitude greater than that of *lacI* (28). (See SI Fig. 6 for systematic analysis of expression levels for local and global TFs.)

Thus far, we have demonstrated that the rapid search hypothesis is biophysically feasible and that colocalization is widespread, even when coregulation effects are excluded, but we have not shown that the selfish gene cluster hypothesis does not explain colocalization. To test this, we considered the relative orientation of a TF gene and the TU it regulates. We compared two TF–TU orientations: downstream unidirectional and convergent (Fig. 4). Other orientations are not considered because they can be due to co/self-regulation (Fig. 3*A*). Although both orientations provide the same recombination distance (i.e., both orientations are equal under the selfish gene cluster hypothesis), rapid search favors the unidirectional orientation because it provides a smaller travel distance (Fig. 4*B*). In case of the convergent orientation, the travel distance is larger because the site is separated from the TF by a whole TU (≈1–5 kbp).

Fig. 4 presents the frequency of TF–TU pairs in these orientations. Remarkably, colocalized TF–TU pairs in the unidirectional orientation are eight times more common than convergent pairs. For a control, we compared TF–TU pairs located at larger distances and observed no difference in frequencies of the two orientations. The strong prevalence of unidirectional arrangements for colocalized TF–TU pairs suggests that the major driving force behind TF-site colocalization cannot be the pressure to form selfish gene clusters, but may be due to the rapid search mechanism. Although it is possible that unidirectional arrangements may be caused by a functional requirement for read-through, we observed the effect at lengths too long for this to be the cause. To test a possibility that the observed prevalence of unidirectional TF–TU pairs is due to a bias toward unidirectional orientation of nearby operons (29), we compared the frequency of unidirectional and convergent orientations for pairs of nearby operons (TU–TU pairs) in EcoCyc (Fig. 4*C* and SI Fig. 7). This control clearly demonstrates that observed 8-fold excess of unidirectional TF–TU pairs cannot be explained by a very mild trend for coalignment of nearby operons. Some examples of colocalized TF–TU pairs are shown in Fig. 5.

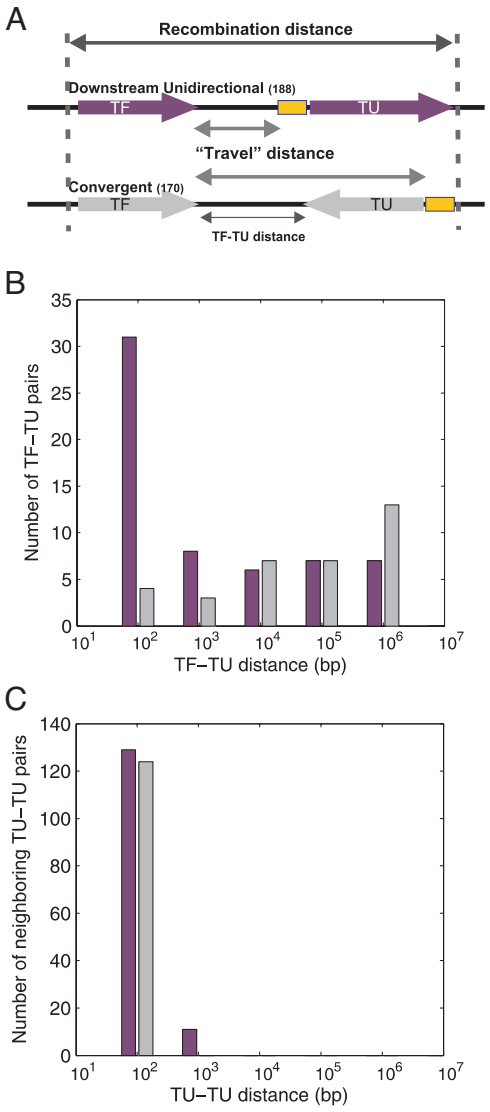

Fig. 4. We test the selfish gene cluster hypothesis as a reason for colocalization. (*A*) Two considered orientations of a TF and its target TU: downstream unidirectional and convergent. Both orientations have the same recombination distance and thus are equally favored by the selfish gene cluster hypothesis. But, only downstream unidirectional orientation provides small travel distance for TF to find its site and therefore is favored by the rapid search hypothesis. Other TF–TU orientations were not considered as they can play a functional role to provide self/coregulation (see Fig. 3*A*). (*B*) The frequency of unidirectional (purple) versus convergent (gray) orientations of TF genes and their target TUs as a function of TF–TU distance. The observed prevalence of unidirectional TF–TU pairs cannot be explained by the selfish gene cluster hypothesis but provides evidence to support the rapid search mechanism. TF–TU distance was measured between the closest ends of the TF and the TU and thus does not depend on orientation or number of ORFs in the TU. Notice that both orientations are equally likely at larger separations (>5 kbp), ruling out an *a priori* bias toward the unidirectional orientation due to other factors. TF–TU pairs at distances <100 bp were neglected to exclude the possibility that TF–TU pairs belonging to the same operon or a read-through locus caused the prevalence of the unidirectional orientation. (*C*) Histogram of the number of TU–TU pair for two downstream TU orientations. This histogram includes data from Ecocyc for verified Escherichia coli Tus (operons). Here, we see that there is no general bias toward the unidirectional configuration in the genome (also see SI Fig. 7).

Discussion

Although our analysis above considers only a subset of TFs, the rapid search hypothesis is quite general. For example, although we excluded from our analysis TFs that are parts of operons or share promoters with their regulated TUs, such gene order is consistent with the rapid search mechanism, because functional organization and biophysical constraints are met simultaneously. Moreover, according to rapid search mechanism, self-regulating operons can benefit from having the TF gene on the first place in the operon (and thus closest to the target promoter). Indeed, we found >3-fold enrichment of TF genes among the first genes in multigene operons (SI Fig. 8).

We also showed that the global (pleiotropic) TFs do not colocalize with their target sites. Clearly, positioning of several regulated operons close to their TF gene is nearly impossible. Pleiotropic TFs are likely to achieve rapid search by being present in high copy number. We also note that, although TF genes and their sites may not be close along DNA, they may be proximal in space because of the organization of DNA in the cell (4, 30) or looping of DNA (31, 32), thus opening a possibility of gene regulation by DNA conformation (33, 34). DNA conformation may also play an important role in the search process (11)

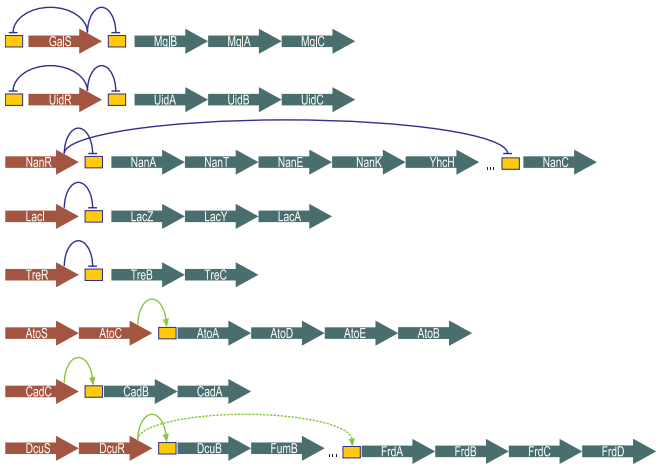

Fig. 5. Examples of colocalized TF–TU pairs. TF genes are shown in red, regulated TU genes in green, negative regulation (repression) as blue blunt arrows, and positive regulation (activation) as green arrows, where dotted arrows represent weak regulation. In the first example, GalS represses the adjacent galactose transport operon *mglBAC* and itself, forming a negative-feedback loop. In the presence of galactose, GalS dissociates from its binding sites at the *mglBAC* and *galS* promoters, starting the transcription of these genes. However, the difference in DNA binding constants between GalS and the GalS–galactose complex is relatively small, on the order of 2 orders of magnitude. Experimental data suggest (45) that even in the presence of galactose, when GalS reaches a sufficiently large concentration, the repressor again shuts down the transcription of transporter genes encoded in *mglBAC* operon. It can be speculated that at the low expression levels typical for local regulators such as GalS and LacI, this is possible only in cases where a high local concentration of GalS is reached, making the release of the newly synthesized GalS protein near the 3′ end of *galS* gene and the 5′ end of *mglBAC* operon a key factor.

because, upon a jump, a TF may associate to DNA in a place that is likely to be proximal along the DNA sequence and still reach the site quickly, effectively increasing the distance that provides faster search up to $\approx 10^3$ to 10^4 bp. This picture is consistent with observed periodicity in the distances between a TF gene and the target sites for pleiotropic TFs (4).

The time it takes a transcription factor to find its binding site is a biologically relevant quantity for both activators and repressors. Prokaryotic activators are often activated by small molecules that diffuse very rapidly through the cell; therefore, the activation of activators is not the rate-limiting step. (Using a very conservative estimate, we find that a small molecule can bind its target protein in <1 sec.) In contrast to many eukaryotic activators, prokaryotic activators also do not reside on the promoters while inactive, waiting for activation. Instead, inactive activators diffuse in the cytoplasm and only upon activation find their cognate sites on DNA (e.g., catabolite activator protein) (20). Therefore, the binding of the activator to its binding site and the subsequent recruitment of RNA polymerase are the rate-limiting steps for the alteration of gene expression.

The search time of repressors for their binding sites is also biologically relevant. In many cases, repressors regulate the production of proteins that are toxic to the cell when produced at inappropriate times. For example, the production of tetracycline resistance operon (35) or lactose permease when it is not needed confers a measurable fitness disadvantage (36). Slow search times lead to leaky repression, which increases the steady-state level of otherwise repressed toxic proteins in the cell.

One surprising result of our study is that the global search by a low-copy-number TF for its site is slow. This result goes against previous estimates for the search time (10, 13, 37, 38) that predominantly used either unrealistically high diffusion coefficients and/or assumed that the fraction of time spent on DNA (or the sliding length) is optimized for fastest search. Our estimate, in contrast, relies on the measured affinity for nonspecific DNA, yielding a much lower rate of binding. As we and others (10, 11) have shown, strong affinity for nonspecific DNA can make search slow, even slower than search by 3D diffusion alone.

Why do TFs have an affinity for nonspecific DNA that makes the search so slow? One possibility is that the affinity for nonspecific DNA is optimized for an equilibrium binding rather than for kinetics. This affinity controls the balance between binding the nonspecific DNA and cognate sites and enables a TF leave its site when the specific affinity to the cognate site drops because of binding of a ligand (20, 38). Our result does not contradict experiments that demonstrate very rapid (faster than 3D diffusion) association of TFs to their sites *in vitro*, because these experiments used concentrations of DNA much lower than that observed in the cell.

Although we have only considered prokaryotes, TFs in eukaryotes also need to rapidly recognize their binding sites. In this case, colocalization will not help because transcription and translation are uncoupled, so they may compensate by (*i*) having a high copy number for global regulators and (*ii*) keeping local TFs constitutively bound to their sites and activating them when necessary [e.g., Gal4 (39)].

Slow spatial diffusion and compartmentalization (40) may favor colocalization in other cellular processes such as signal transduction (see ref. 41 for review) or interactions between receptors on the membrane (42).

In summary, we used simulations to show that the colocalization of a TF gene and its sites is required for rapid, reliable regulation of gene expression by low-copy-number TFs. We demonstrated that widespread colocalization of local TFs and their targets in bacterial genomes exists and cannot be fully attributed to co/self-regulation or the selfish gene cluster hypothesis. We conclude that rapid and reliable gene regulation imposes a biophysical constraint on the organization of bacterial genomes, encouraging TF genes and their binding sites to be close.

Materials and Methods

Simulating a Transcription Factor's Search for Its Binding Site. To explore the kinetic effects of TF–TU gene colocalization, we simulated a transcription factor's search for its binding site and varied the starting position of the TF. We modeled a typical prokaryotic genome as a string 10^7 bp and randomly selected a binding site. We placed the TF at a given distance along the chromosome from the binding site and then simulated alternating rounds of 3D diffusion and 1D sliding until the transcription factor found its binding site. Sliding along the chromosome was modeled as an explicit 1D random walk. We simulated 3D diffusion as a mixture of hops, short correlated motions through the cell volume, and jumps, long, uncorrelated movements. The details of the simulation are described in the *SI Text* and SI Table 1.

Data Acquisition and Preparation. LacI family members were identified by using several databases and algorithms (*SI Text*). The SignalX program (43) was used to identify the binding motifs for TFs and construct the recognition profiles. Candidate sites were identified by scanning the genomes with the constructed profiles. Only orthologous binding sites, that is, binding sites occurring upstream of orthologous operons were retained for further analysis. This resulted in identification of 159 TFs and 647 binding sites from 36 genomes. These data are deposited in the RegTransBase database (http://regtransbase.lbl.gov). A summary of the data are presented in SI Table 2.

Because of the reliability of the data, here, we present our analysis of the LacI data set. However, we carried out a similar analysis using the EcoCyc data set (44), which provides more

complete, if slightly less reliable, TF–TU data, and the results are presented in SI Figs. 9 and 10.

We defined several classes of transcription factors and binding sites. The global set includes the pleiotropic TFs (FruR, PurR, and CcpA), which each bind more than four sites on the genome, and their binding sites (25, 26), and the local set includes all of the nonpleiotropic TFs and their binding sites. To avoid the strong but unrelated signal generated by self-regulating TFs, we excluded binding sites residing within the 5′ operator region of the corresponding TF gene (Fig. 3*A*).

Measuring Distances Between Genetic Objects. The distance between two genetic objects was measured in base pairs and was defined as the distance between the two nearest nucleotides of the objects, regardless of the direction. In this article, we use TF–TU distance, the distance between a TF and the nearest regulated TU gene.

We thank Mehran Kardar, Johnathan Widom, Shamil Sunyaev, Hanah Margalit, Nir Fridman, Ido Golding, and Alexander Grosberg for useful discussions. L.A.M. and G.K. are supported by the National Center for Biomedical Computing, i2b2. O.N.L. and M.S.G. are partially supported by International Association for the Promotion of Cooperation with Scientists from the New Independent States of the Former Soviet Union (INTAS) Grant 05-1000008-8028 and the Russian Academy of Sciences (program "Molecular and Cellular Biology"). M.S.G. is a Howard Hughes Medical Institute International Research Scholar. Z.W. is a Howard Hughes Medical Institute Predoctoral Fellow.

1. Pardee AB, Jacob F, Monod J (1959) *J Mol Biol* 1:165–178.
2. Warren PB, ten Wolde PR (2004) *J Mol Biol* 342:1379–1390.
3. Korbel JO, Jensen LJ, von Mering C, Bork P (2004) *Nat Biotechnol* 22:911–917.
4. Kepes F (2004) *J Mol Biol* 340:957–964.
5. Hershberg R, Yeger-Lotem E, Margalit H (2005) *Trends Genet* 21:138–142.
6. McFall E (1986) *J Bacteriol* 167:429–432.
7. Golding I, Cox EC (2006) *Phys Rev Lett* 96:098102.
8. Widom J (2005) *Proc Natl Acad Sci USA* 102:16909–16910.
9. Berg OG, Winter RB, von Hippel PH (1981) *Biochemistry* 20:6929–6948.
10. Slutsky M, Mirny LA (2004) *Biophys J* 87:4021–4035.
11. Hu T, Grosberg AY, Shklovskii BI (2006) *Biophys J* 90:2731–2744.
12. Coppey M, Benichou O, Voituriez R, Moreau M (2004) *Biophys J* 87:1640–1649.
13. Halford SE, Marko JF (2004) *Nucleic Acids Res* 32:3040–3052.
14. Redner S (2001) *A Guide to First-Passage Processes* (Cambridge Univ Press, Cambridge, UK).
15. Lawrence JG, Roth JR (1996) *Genetics* 143:1843–1860.
16. Lawrence J (1999) *Curr Opin Genet Dev* 9:642–648.
17. Gowers DM, Wilson GG, Halford SE (2005) *Proc Natl Acad Sci USA* 102:15883–15888.
18. Smoluchowski MV (1917) *Z Phys Chem* 92:129–198.
19. Revzin A (1990) *The Biology of Nonspecific DNA Protein Interactions* (CRC, London).
20. Ptashne M (1992) *A Genetic Switch* (Cell, Cambridge, MA).
21. Elf J, Li GW, Xie XS (2007) *Science* 316:1191–1194.
22. Link AJ, Robison K, Church GM (1997) *Electrophoresis* 18:1259–1313.
23. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) *Cell* 123:1025–1036.
24. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) *Science* 311:1600–1603.
25. Martinez-Antonio A, Collado-Vides J (2003) *Curr Opin Microbiol* 6:482–489.
26. Tobisch S, Zuhlke D, Bernhardt J, Stulke J, Hecker M (1999) *J Bacteriol* 181:6996–7004.
27. Tan K, McCue LA, Stormo GD (2005) *Genome Res* 15:312–320.
28. Allen TE, Herrgard MJ, Liu M, Qiu Y, Glasner JD, Blattner FR, Palsson BO (2003) *J Bacteriol* 185:6392–6399.
29. Price MN, Huang KH, Alm EJ, Arkin AP (2005) *Nucleic Acids Res* 33:880–892.
30. Wright MA, Kharchenko P, Church GM, Segrè D (2007) *Proc Natl Acad Sci USA* 104:10559–10564.
31. Gowers DM, Halford SE (2003) *EMBO J* 22:1410–1418.
32. Vilar JM, Saiz L (2005) *Curr Opin Genet Dev* 15:136–144.
33. Jeong KS, Ahn J, Khodursky AB (2004) *Genome Biol* 5:R86.
34. Peter BJ, Arsuaga J, Breier AM, Khodursky AB, Brown PO, Cozzarelli NR (2004) *Genome Biol* 5:R87.
35. Lenski RE, Souza V, Duong LP, Phan QG, Nguyen TN, Bertrand KP (1994) *Mol Ecol* 3:127–135.
36. Dykhuizen D, Hartl D (1978) *J Bacteriol* 135:876–882.
37. Winter RB, Berg OG, von Hippel PH (1981) *Biochemistry* 20:6961–6977.
38. Gerland U, Moroz JD, Hwa T (2002) *Proc Natl Acad Sci USA* 99:12015–12020.
39. Selleck SB, Majors JE (1987) *Mol Cell Biol* 7:3260–3267.
40. Bork P, Serrano L (2005) *Cell* 121:507–509.
41. Kholodenko BN (2006) *Nat Rev Mol Cell Biol* 7:165–176.
42. Batada NN, Shepp LA, Siegmund DO, Levitt M (2006) *PLoS Comput Biol* 2:e44.
43. Mironov AA, Vinokurova NP, Gel'fand MS (2000) *Mol Biol (Moscow)* 34:253–262.
44. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) *Nucleic Acids Res* 33:D334–D337.
45. Geanacopoulos M, Adhya S (1997) *J Bacteriol* 179:228–234.
46. Wunderlich Z, Mirny LA (2007) *Spatial Effects on the Speed and Reliability of Protein–DNA Search*, www.arxiv.org/abs/0708.1136.