Computational and Structural Biotechnology Journal 10 (2014) 63–69

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression

Daphne Ezer [*,1], Nicolae Radu Zabet [*,1], Boris Adryan [*]

Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK
Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

ARTICLE INFO

Available online 1 August 2014

Keywords:
Facilitated diffusion
Cooperativity
Massively parallel gene expression assays
Enhancers
Mechanistic models

ABSTRACT

The organization of binding sites in cis-regulatory elements (CREs) can influence gene expression through a combination of physical mechanisms, ranging from direct interactions between TF molecules to DNA looping and transient chromatin interactions. The study of simple and common building blocks in promoters and other CREs allows us to dissect how all of these mechanisms work together. Many adjacent TF binding sites for the same TF species form homotypic clusters, and these CRE architecture building blocks serve as a prime candidate for understanding interacting transcriptional mechanisms. Homotypic clusters are prevalent in both bacterial and eukaryotic genomes, and are present in both promoters as well as more distal enhancer/silencer elements. Here, we review previous theoretical and experimental studies that show how the complexity (number of binding sites) and spatial organization (distance between sites and overall distance from transcription start sites) of homotypic clusters influence gene expression. In particular, we describe how homotypic clusters modulate the temporal dynamics of TF binding, a mechanism that can affect gene expression, but which has not yet been sufficiently characterized. We propose further experiments on homotypic clusters that would be useful in developing mechanistic models of gene expression.

Contents

1. Introduction

Gene expression is largely determined by the combination of transcription factors (TFs) that are bound to promoters or other cis-regulatory elements (CREs, also known as CRMs — cis-regulatory modules). These cis-regulatory regions contain multiple closely spaced and sometimes overlapping binding sites [1–3]. Simple models of gene

[*] Corresponding authors.
E-mail addresses: de276@cam.ac.uk (D. Ezer), n.r.zabet@gen.cam.ac.uk (N.R. Zabet), ba255@cam.ac.uk (B. Adryan).
[1] These authors contributed equally to the paper as first authors.

http://dx.doi.org/10.1016/j.csbj.2014.07.005
2001-0370/© 2014 Ezer et al. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

expression often consider how each TF contributes individually to gene expression [4–6], but recent synthetic biology studies have demonstrated that also the order and spacing of binding sites can influence transcription rates [7–10]. Thus, there is a need to move towards more detailed models of gene regulation that consider the regulatory sequence in conjunction with the abundance of TFs and the dynamical behavior of the system [11].

Even though the order, orientation and spacing of binding sites have received some attention and have been the focus of experimental studies, the field is far away from a true mechanistic and predictive model of how DNA sequence encodes regulatory information. Instead we are currently presented with a range of possible mechanisms that explain how promoter organization can influence transcription, such as: (i) the dynamics of TF binding [1,12], (ii) nucleosome displacement [13,14], (iii) protein–protein interactions [15–17] and (iv) DNA looping and TF interactions with the transcriptional machinery (such as the mediator complex) [18–20].

TF binding sites can be organized in many combinations across the genome; so we are left with the difficult task of finding out how these diverse TF binding site architectures influence the physical mechanisms that ultimately lead to transcription. A first step towards developing a more mechanistic view of CRE organization is to dissect common and simple organizational patterns [1]. One of the most common CRE building blocks is the *homotypic cluster*, a group of adjacent binding sites for the *same* TF. They are found in bacterial and eukaryotic promoters, as well as in eukaryotic CREs.

In this review, we will argue that homotypic clusters can serve as an excellent model system for understanding how complex physical processes interact to control gene expression. We present several examples of homotypic clusters and propose distinguishing characteristics of their potential mechanisms. Finally, we provide biological examples of homotypic clusters in several organisms, ranging from bacteria [1,2] to fruit fly [21] and mammalian genomes [22–24], which illustrates the importance of homotypic clusters in biological systems.

2. Homotypic clusters as a model system for studying complex CREs

Recently, it has become possible to synthesize thousands of promoters or enhancers, and to measure the resulting level of gene expression in parallel, an experimental design known as a massively parallel gene expression assay [9,25,26]. With this new technology, it is possible to experimentally test how different TF binding site organizations influence gene expression.

Even with the development of techniques to synthesize DNA more efficiently, it is still very difficult to study how heterotypic clusters influence gene expression. As the number of adjacent TF binding sites increases, the number of possible permutations of binding sites expands at a factorial scale. The distance between the binding sites and the order of the binding sites may also influence the TF–TF interactions, further increasing the total number of possible binding site organizations that would need to be systematically assessed for a complete characterization. Smith et al. [10] randomly sampled a subset of these TF binding site permutations in a massively parallel gene expression assay. Different permutations produced significantly different levels of transcription, but their approach was unable to identify predictive patterns for gene expression.

Homotypic clusters can be used to study the effects of binding site strength, orientation, and positioning, while ignoring the effects of heterotypic TF–TF interactions, drastically reducing the scale of the problem. However, in some cases, a TF is only able to activate transcription in the presence of a co-activator, which is an important consideration when designing synthetic constructs for massively parallel gene expression assays. Smith et al. [10] also conducted these experiments with different sizes of homotypic clusters, but mostly found weak correlations between the size of a homotypic CRE and the resulting level of gene expression. One possible explanation for these weak correlations is that their experimental design did not include potentially essential cooperative proteins [10].

Another massively parallel gene expression assay indicated that only homotypic clusters that appear to be bound by TFs (as per a ChIP-seq experiment) influenced gene expression [27], suggesting that ChIP experiments should be an integral part of the workflow in these massively parallel gene expression assay experiments for added interpretability. Local sequence context, such as the GC content of sequences flanking the binding site, could drastically influence the binding of TFs in ways that we cannot fully predict [27].

Homotypic clusters are a nearly ubiquitous feature in regulatory regions of organisms ranging from *Escherichia coli* K12 [28] to *Drosophila melanogaster* [29] to humans [22]. Therefore, understanding how homotypic clusters can influence gene expression would provide insight into an important regulatory mechanism in many if not most organisms.

First, we will review the fundamental properties of these regions in terms of their prevalence, sequence conservation, and possible species-specific functional roles.

3. Mechanisms by which homotypic clusters could influence gene expression

There are many physical mechanisms that influence gene expression: from the combination of TFs that are bound, to the chromatin state and to the interaction of TFs with the transcriptional machinery. In what follows, we will systematically review different mechanisms by which homotypic clusters might influence gene expression.

3.1. Assuming no cooperativity

We will start by considering the case of TFs that do not interact with one another at all, and describe how clusters can provide a mechanism for gene regulation even under this simple scenario, which is sometimes called the “billboard model” [5]. Under such a model, each binding site in a cluster has a uniform probability of being bound and this probability may be associated with an external variable to the system, such as TF concentration.

The effect of a homotypic cluster on gene expression under such a model depends on how the TF binding pattern influences gene expression, of which we will consider four cases: (i) all the binding sites must be bound for the gene to be regulated, (ii) at least one binding site must be bound for the gene to be regulated, (iii) each binding site independently contributes to gene expression and (iv) each binding site has a different, but independent, contribution to gene expression (dependent on a property such as distance from the TSS); see Fig. 1.

The first case (all the binding sites must be bound for the gene to be regulated) results in a switch-like behavior of transcription [30,31] and consequently reduces leaky gene expression and noise in mRNA levels [32]. In this scenario, the cluster is acting as a buffer that prevents spurious transcription until the concentration of TF is high enough such that all binding sites are occupied. In addition, such a system generates a time delay in gene regulation; the more binding sites in a cluster, the longer it would take for all the binding sites to be bound [32].

In the second scenario, only a single TF must be bound for transcription to take place, so having long homotypic clusters increases the likelihood of transcription compared to a single site, the opposite of the previous scenario. In this case, homotypic clusters make a promoter more sensitive to low concentrations of TFs and less sensitive to higher concentrations of TFs. In addition, assuming that at least one binding site must be bound in a homotypic cluster decreases at the time it would take for the gene to be regulated [1].

Note that these first two scenarios correspond to “AND logic” (case 1; multiple TFs must be bound to their binding sites) and “OR logic” (case 2; at least one TF must be bound to their binding sites), both of which have been identified as the regulatory logic defining

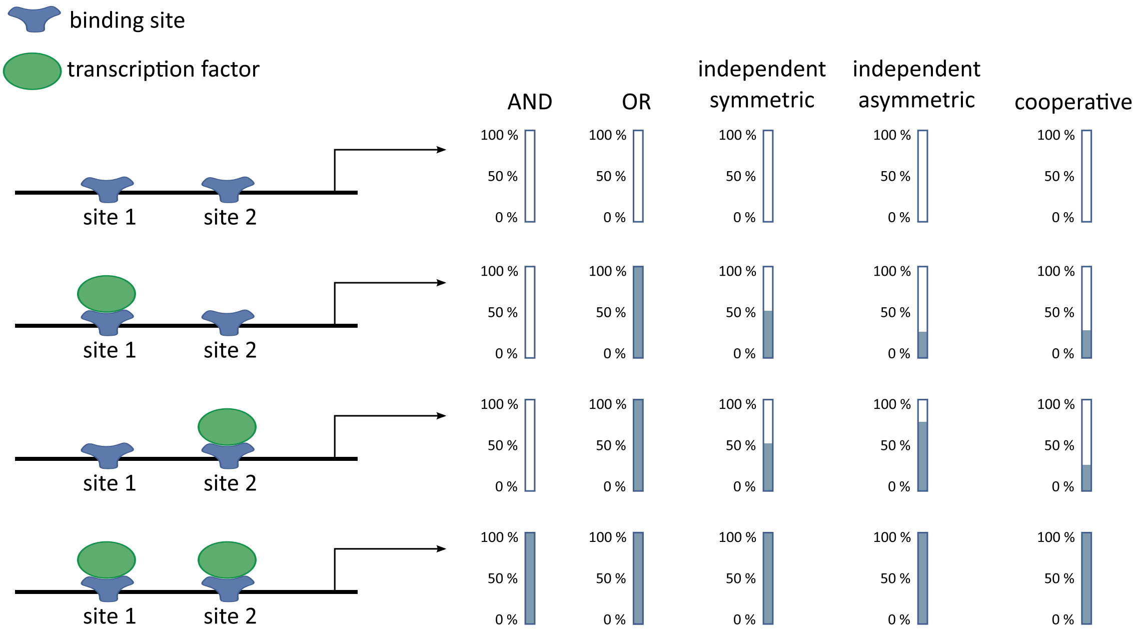

Fig. 1. Binding configurations in a homotypic cluster of two binding sites. To the right, we present the level of gene expression, given various different mechanisms of TF action. The mechanisms of TF action we include is AND logic (all the TFs must be bound for transcription to occur), OR logic (at least one TF must be bound for transcription to occur), independent symmetric (each TF independently contributes to gene expression), independent asymmetric (each TF independently contributes to gene expression, but the effect on transcription is also dependent on the position of the binding site) and cooperative (there may be some leaky expression when each TF independently binds, but there are synergistic effects when both TFs are bound).

both real and synthetic biology systems [33,34] although not specifically in the case of homotypic clusters.

In the third scenario each binding site independently contributes to gene expression; in other words, a cell might have different levels of gene expression, dependent on the number of TFs that are bound. This means that the number of occupied binding sites is correlated with the expression level of a gene. To distinguish between these three scenarios (AND, OR and independent), it is possible to take clusters of binding sites of different lengths and vary the concentration of TFs, measuring gene expression (ideally in a single-cell framework to more clearly distinguish the latter two cases OR and independent). Giorgetti et al. [35] compared these three models for the NF-kappaB system and discovered that this third model explains the experimental data best.

Even though the third scenario seems to be a sufficient model for explaining some real biological systems with homotypic clusters, this model assumes that all the TFs bound to a homotypic cluster contributed equally to expression, which might not always be the case. Certain TFs have optimal distances from the TSS that maximizes their interaction with the transcriptional machinery [7,36]. Alternatively, in some cases, there is a periodic relation between the distance of a TF binding site from the TSS and the level of transcription, possibly because the influence of TFs on gene expression is dependent on the nucleosome context [9].

In conclusion, homotypic clusters can generate a wide range of behaviors, even if we assume that TFs are not cooperative.

3.2. Assuming direct cooperativity

In addition, homotypic clusters can influence gene regulation through direct, physical TF–TF interactions [37]; see Fig. 2. One particular example of TF–TF interaction is the phenomenon of homodimerization, where pairs of molecules of same TF bind directly to each other before binding to DNA [38]. A homodimer has two identical DNA binding domains, usually in opposite orientations. Therefore, a strong indication that a TF forms homodimers is the presence of many binding site pairs in alternating orientations, with a fixed distance between the sites [10,39].

In the case of indirect interactions, Giorgetti et al. [35] saw that TF concentration could have a gradual effect on gene expression. In contrast, if the likelihood of a TF being bound increases with the number of TFs already bound, then one would expect that the number of bound molecules would match a sigmoid curve; see Fig. 2. In other words, homotypic clusters without TF–TF interaction would result in analog regulatory logic, while homotypic clusters with TF–TF interactions would result in digital regulatory logic [30,35]; see Fig. 2.

In *D. melanogaster*, many of the homotypic clusters are found in developmental genes that require such a binary behavior, and protein–protein interactions have been proposed as playing a role in achieving this [21,35]. In fact, bicoid, one of the primary TFs that form the main anterior–posterior axis in the early embryo, likely operates in this way [40].

Another advantage of homotypic clustering with direct TF–TF interaction is increased binding stability. In mammals, highly degenerate TF binding sites that are conserved tend to occur in homotypic clusters [24]. Possibly, these binding sites are not strong enough to bind TFs individually, but the TFs can stabilize each other's binding.

3.3. Assuming indirect cooperativity

Even if two proteins do not physically interact with one another, they can affect each other's ability to bind [41]. For instance, some models assume that proteins bind to the genome at thermodynamic equilibrium [42] and, in this scenario, the presence of many weak TF binding sites might result in nucleosome displacement being the most energetically favorable conformation [13,14]. Therefore, homotypic clusters may allow TFs to stabilize each other's binding, even without direct TF–TF interaction.

Another case of indirect cooperativity is the effect of binding site co-localization on the binding/unbinding kinetics of TFs from their binding sites. Riggs et al. [43] observed that lac repressor (a bacterial TF) binds 100–1000 times faster to its target site than would be possible by simple three-dimensional diffusion alone. It seems that when binding to their target sites, TFs perform a combination of three-dimensional diffusion in the cytoplasm/nucleoplasm and one-dimensional random walk on

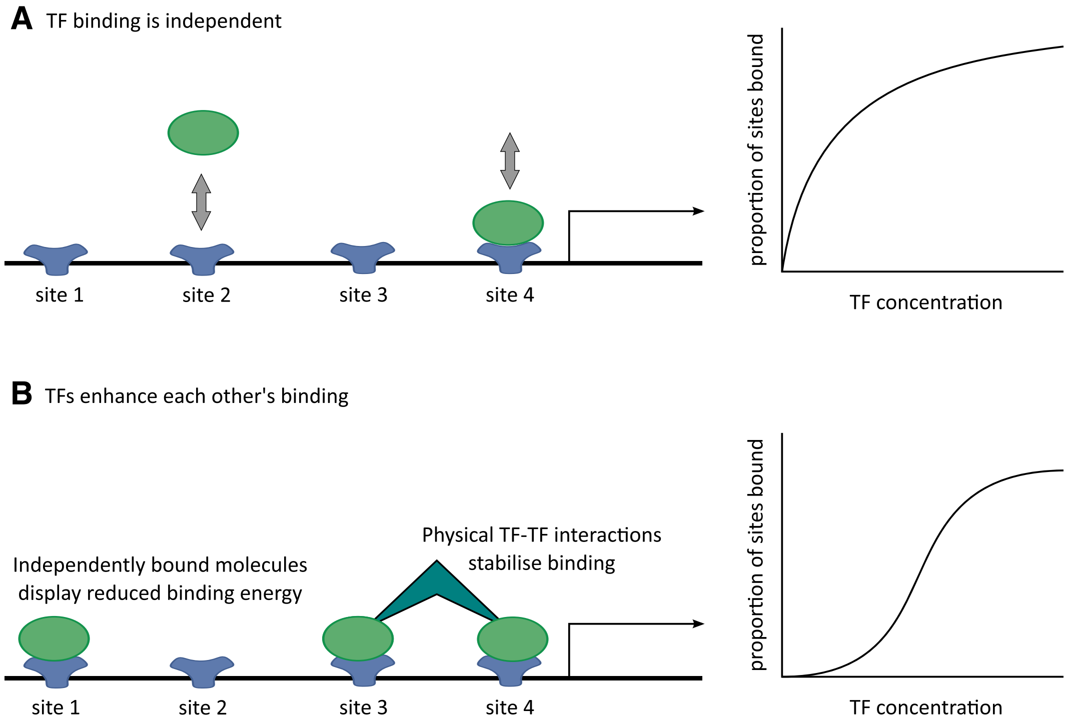
A TF binding is independent
site 1
site 2
site 3
site 4
proportion of sites bound
TF concentration
B TFs enhance each other's binding
Independently bound molecules display reduced binding energy
Physical TF-TF interactions stabilise binding
site 1
site 2
site 3
site 4
proportion of sites bound
TF concentration

Fig. 2. Cooperative binding assuming direct TF–TF interactions. We illustrate the cases of TF binding independently to their binding sites (A) and cooperatively through direct TF–TF interactions (B). (A) We assumed that the binding of TF molecules to the binding sites is independent and in this case the proportion of bound sites increases gradually with the TF concentration. (B) We assumed that direct TF–TF interaction can stabilize the binding and, in this scenario, the proportion of bound sites as a function of TF concentration displays a sigmoid shape. Note that on the right side we plot the proportion of sites bound in each case as a function of TF concentration.

the DNA; reviewed in [44]. This mechanism is known as the facilitated diffusion mechanism and was first formalized by Berg and co-workers [45,46]. The speedup in the search process is a consequence of the reduction of the dimensionality of the search process from three-dimensions to one-dimension. Despite errors in the original calculations [47], later studies provided experimental evidence of the existence of this mechanism. One indirect piece of evidence for the existence of facilitated diffusion is that TFs have a higher association rate in vitro to longer synthesized DNA fragments compared to shorter DNA fragments despite the fact that both longer and shorter DNA fragments contain the same binding site in the middle [48]. This mechanism is called the antenna effect and assumes that having a longer DNA fragment increases the contribution of the one-dimensional random walk component to the TF search process [49]. The most conclusive evidence comes from direct observation of the movement of DNA binding molecules, which were performed in vitro [50–52] and in vivo [12,53–56]. In particular, the first experimental evidence of the existence of the one-dimensional random walk on the DNA was provided by Kabata et al. [50], who observed linear movements of *E. coli* RNAp on the DNA. However, a recent study performed by Wang et al. [57] visualized *E. coli* RNAp diffusion in vitro and found that the RNAp mainly performs three-dimensional diffusion, while the contribution of one-dimensional random walk to the search process is marginal. While we know that some DNA binding proteins spend more time performing the one-dimensional random walk on the DNA compared to others, we still do not understand what determines this preference.

In the context of facilitated diffusion, binding site co-localization could lead to effects that cannot be captured by statistical thermodynamics models. In [1], we proposed that the co-localization of binding sites can be decomposed into one of three building blocks: (i) switches (overlapping sites), (ii) barriers (closely spaced sites) and (iii) homotypic clusters. In the former, only one TF can be bound at once due to steric hindrance, resulting in switch-like behavior [58]. In the second case, the presence of a nearby site can reduce the association rate of a TF to its target site by blocking the one-dimensional search from one direction; the so called barrier effect [1,12,59]. Lastly, we identified that a homotypic cluster can have a dual role: (i) it can result in keeping the TF molecule longer within one region by sampling several high affinity sites during a single one-dimensional random walk on the DNA and (ii) it can result in a barrier effect [1]; see Fig. 3. In other words, there is a tradeoff related to the optimal spacing between binding sites: large spaces between TFs in a homotypic cluster would decrease the time for the second binding site to be occupied, smaller spaces would increase the time a TF molecule will spend in that region. A recent study also showed experimentally that, in the context of facilitated diffusion, homotypic clusters do not only seem to affect mean expression levels, but also the noise in gene expression [60].

We have previously found that, in combination with other promoter organizational motifs, homotypic clusters can generate complex binding dynamics over time. For example, the occupancy of a binding site flanked by two homotypic clusters displays an impulse, with a fast increase in occupancy and then a decrease to a lower level [1]. It should be noted that this complex promoter organization (binding site flanked by two homotypic clusters) is encountered seven times in the *E. coli* genome [1].

Additionally, the presence of weaker sites flanking a strong binding site could lead to a funnel effect where the molecules are directed to the strong binding site and retained there for longer times [61–63]. However, it is controversial whether TFs can bind at all to weak binding sites [64], which is a necessary assumption required for the funnel effect hypothesis. Nevertheless, experiments that verify the facilitated diffusion mechanism have focused on only a handful of TFs, so it is uncertain whether all TFs perform facilitated diffusion. In addition, it is extremely difficult to demonstrate whether facilitated diffusion influences binding dynamics in a biologically significant way in vivo.

4. Bacteria

Bacteria regulatory regions are usually condensed regions of a few hundred base pairs immediately adjacent to the transcription start

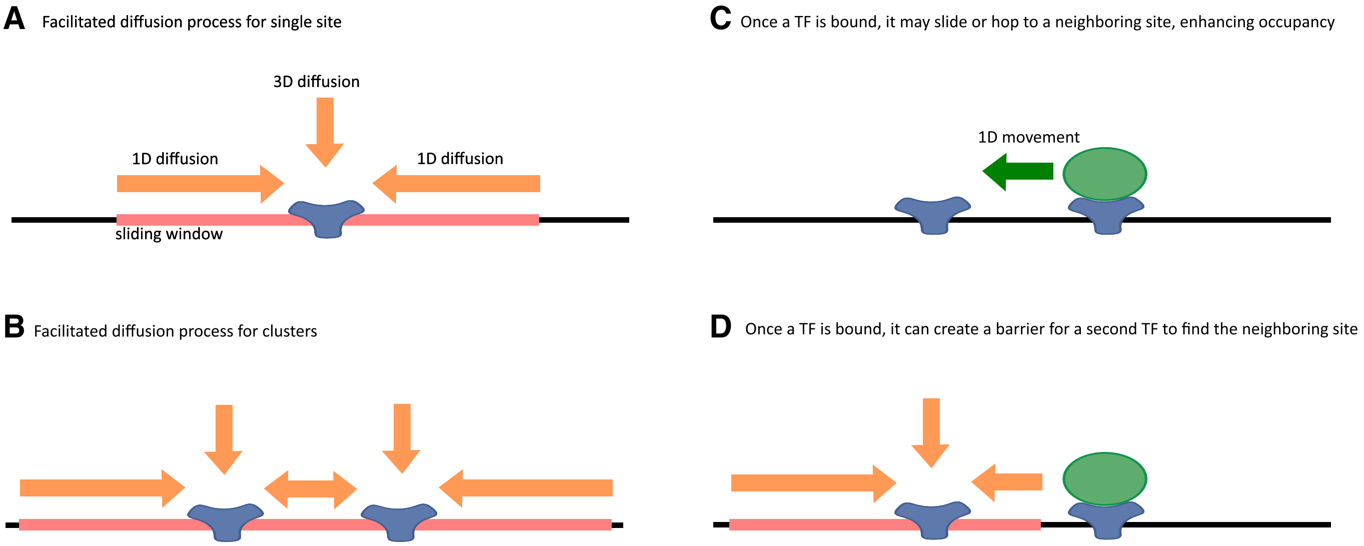
A Facilitated diffusion process for single site
3D diffusion
1D diffusion
1D diffusion
sliding window
B Facilitated diffusion process for clusters
C Once a TF is bound, it may slide or hop to a neighboring site, enhancing occupancy
1D movement
D Once a TF is bound, it can create a barrier for a second TF to find the neighboring site

Fig. 3. The influence of homotypic clusters on facilitated diffusion. (A) We illustrate the process of facilitated diffusion for a single binding site: a TF can find the binding site by 3D diffusion or 1D diffusion from either side (orange arrows). If a TF randomly binds within the sliding window (illustrated by the red bar) then it will diffuse to the binding site with high probability. Therefore, the longer the red bar, the higher the probability a TF will find its binding site. In panels (B–D), we illustrate three ways in which homotypic clusters influence TF search time and occupancy within a facilitated diffusion context. (B) The sliding window is expanded by the presence of the homotypic clusters, so it is faster to find a site in a cluster than an individual site. (C) Once a TF is bound, it may slide or hop randomly to the neighboring binding site, thereby enhancing TF occupancy. (D) If one TF is already bound to a binding site, then it restricts the sliding length by which a second TF can find its binding site, which we refer to as the barrier effect. Therefore, homotypic clusters influence temporal dynamics of TF binding in a variety of ways.

site. Our previous work cataloged all promoter organizations in *E. coli* K12, including those promoter organizations that incorporate homotypic clusters [1] (see http://logic.sysbiol.cam.ac.uk/fgrip/db/) and showed that homotypic clusters are one of the most prevalent organizations of 3–5 closely spaced binding sites. In particular, we found eleven promoters with at least five immediately adjacent binding site repetitions (dadAp1, frdAp, metFp, metBp, csgBp, entCp, fepAp, acnBp, pdhRp, proVp3, narKp1), most of which include overlapping binding sites. Within a facilitated diffusion context, a TF might hop or slide between neighboring sites (thus, enhancing occupancy) and the closer the two neighboring binding sites, the stronger these effects would be.

Most of the TFs in these long clusters have very frequently occurring binding sites and also have high cellular concentrations (e.g. H-NS, NarL, CRP, ArcA, Fur, CpxR, MetJ, LRP) [1,28]. Since the TF concentrations are so high, we would expect that these TFs would find their binding sites quickly. However, many of these TFs also have binding motifs that are not very specific, so having many repeating binding sites could help maintain a higher local concentration of TFs near the DNA regions where they provide important regulatory functions.

5. Eukaryotes

In mammals, homotypic clusters occupy a large portion of the genome. For example, in humans, homotypic clusters cover approximately 1.6% of the genome (on the same order of magnitude as exons) and almost half of the 487 experimentally verified CREs have homotypic clusters. In addition, the binding sites in the homotypic clusters are more conserved than the space between the sites, with the central TF binding site often the most conserved, suggesting that the binding sites may be under a purifying selection [22]. These homotypic clusters are also enriched in proximal enhancers and promoters, particularly in bidirectional promoters, and they are often associated with the Ep300 protein (also known as p300), which is associated with the mediator complex [22]. Taken together, these results indicate that many homotypic clusters are probably associated with active genes [22] and that homotypic clusters are prevalent and likely to be functionally important.

In particular, among mammals, the two most significant GO terms associated with homotypic clusters are “protein binding” and “transcription factor activity” (other enriched GO terms include “nucleotide binding”, “sequence specific DNA binding” and “regulation of transcription, DNA binding”) [22]. Overall, 62% of annotated TF genes have homotypic clusters within their promoters. Among homotypic clusters that are conserved across vertebrates (frog, chicken, mouse, and human), there is even greater enrichment for homotypic clusters in the promoters of genes encoding TFs [22]. In fruit flies, they are found in many known developmental CREs, with many key developmental regulators such as Bicoid and Kruppel forming clusters [21]. A small change in the transcription rate of a TF (especially a TF involved in early development) might influence the transcription rates of a large number of downstream genes; therefore, the cell must carefully modulate the transcription rates of these genes. Depending on whether or not the TFs cooperate with one another, homotypic clusters could allow gene networks to provide analog or digital responses to changing concentrations of TFs, influencing their fundamental dynamics.

The mammalian genome is scattered with highly degenerate TF binding sites, DNA sequences that are slightly similar to TF binding motifs, but that do not constitute a significant match to a TF binding motif as suggested by their computed affinity. Some studies suggest that TFs might not recognize these weak binding sites [64] and that TFs may only influence transcription if they are bound to strong sites [65]. Surprisingly, these highly degenerate TF binding sites are likely to be part of homotypic clusters, and those degenerate sites that form clusters are often significantly conserved across mammalian species [24], suggesting that these degenerate binding sites may be functional after-all. One potential explanation for the high conservation of the homotypic clusters is that TF–TF cooperation within the homotypic clusters might stabilize binding to weak sites or that homotypic clusters enhance the local concentration of TFs in these regions.

6. Prevalence versus functional significance

Evidence from multiple sources indicates that the presence of homotypic clusters can result in certain patterns in gene expression; however, there might be simpler ways to obtain these behaviors. For instance, instead of having many weak binding sites for a TF, why is there not a single strong and stable binding site?

Hermsen et al. [66] found that by simulating promoter evolution of simple promoters to optimize for certain types of transcriptional logic, homotypic clusters would often emerge, indicating that homotypic clusters may in fact be the easiest-to-achieve solution to certain selective pressures for some regulatory logic patterns. On the other hand, some researchers have argued that the abundance of homotypic clusters in the genome of so many organisms may not be caused by the evolutionary advantages of this organization. Rather, the way mutations accumulate (the so called *sampling of the genotypic–phenotypic landscape*) would result in the enrichment of homotypic clusters [67,68]. For instance, there may be many ways to reach a certain level of gene expression, but a high proportion of possible solutions include homotypic clusters and therefore they appear relatively frequently [68].

One explanation for the high frequency of homotypic clusters is that homotypic clusters lie in a "flatter" portion of the genotype–phenotype landscape and, thus, mutations are less likely to affect the function of a homotypic cluster than they would for a single strong binding site [69]. In the latter case, natural selection is acting on "ability to withstand mutations" rather than "phenotypic optimality".

Given the low DNA specificity for eukaryotic TFs [70], spontaneous homotypic binding sites can arise by chance alone. In particular, some homotypic clusters occur in regions of short tandem repeats, which can change their size rapidly in just a few generations due to the process of DNA slippage [71]. Nevertheless, previous studies showed that, for some multicellular eukaryotes, the homotypic cluster formation within short distances (50 bp) is most likely a consequence of local sequence duplication than of point mutations, while, in the case of bacteria or unicellular eukaryotes, point mutations are most likely to be the source of homotypic clusters [72].

Therefore, the prevalence of specific promoter architectures in the genome does not indicate that it is important for gene regulation. Nevertheless, massively parallel gene expression assays demonstrate that manipulating the properties of homotypic clusters can influence gene expression and noise [9,10,60]. In addition, homotypic clusters are often conserved across divergent species [22]. This suggests that homotypic clusters affect gene expression and are under purifying selection. In conclusion, the model that the evolution of homotypic clusters occurs because of random sampling of the genotypic–phenotypic landscape does not account for all of the observations, although it serves as important null hypothesis [68].

7. Conclusions and outlook

Homotypic clusters are commonly found in organisms ranging from bacteria to humans. Several mechanisms by which homotypic clusters could influence transcription rates have been proposed, but despite the fact that homotypic clusters are the simplest examples of organizational patterns in CREs, they are not well understood. In this review, we presented three cases where homotypic clusters influence gene regulation, namely: (i) when there are no cooperative interactions, but the activity state of the gene is a function of the number of sites occupied in the homotypic cluster; (ii) when there is direct TF–TF interaction and the homotypic clusters allow the binding of the oligomers; (iii) the co-localization of binding sites affect the binding/unbinding dynamics of TFs. While the first two receive significant attention from the literature, the latter case is often neglected. In a recent study we showed how the co-localization of binding sites affects the binding/unbinding kinetics and the occupancy of the binding sites [1]. These results are supported by previous experimental studies, which showed that the presence of "road blocks" on the DNA seems to significantly affect the association rate [12]. Nevertheless, a systematic experimental analysis is still required in order to decompose the contribution of each of these mechanisms to the gene regulation process. One experiment that is essential in generating a comprehensive picture of the role of homotypic clusters on gene regulations consists of comparing the effects of different cluster sizes and different distances between binding sites.

The lac repressor system is a well studied system that would make a good candidate for this analysis [12]. One disadvantage of using this system is that lacI stays bound to its target site for 5 min [73]. Eukaryotic TFs are bound for less time at specific sites (10–20 s) [54,55,74], which potentially makes the results of the lacI system valid only in the context of bacterial cells. Thus, one should design an experiment in a eukaryotic system and, given the current development of precise genome editing tools such as CRISPR/CAS9 system [75], we hope that such data will become available at some stage in the near future.

Funding

This work was supported by the Medical Research Council [G1002110].

Acknowledgment

We would like to thank X. Ma for useful discussions and comments on the manuscript.

References

[1] Ezer D, Zabet NR, Adryan B. Physical constraints determine the logic of bacterial promoter architectures. Nucleic Acids Res 2014;42(7):4196–207. http://dx.doi.org/10.1093/nar/gku078.

[2] Hermsen R, Tans S, ten Wolde PR. Transcriptional regulation by competing transcription factor modules. PLoS Comput Biol 2006;2(12):e164. http://dx.doi.org/10.1371/journal.pcbi.0020164.

[3] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013;153(2):307–19. http://dx.doi.org/10.1016/j.cell.2013.03.035.

[4] Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Res 2012;40(2):553–68. http://dx.doi.org/10.1093/nar/gkr752.

[5] Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 2012;13(9):613–26. http://dx.doi.org/10.1038/nrg3207.

[6] Ilsley GR, Fisher J, Apweiler R, DePace AH, Luscombe NM. Cellular resolution models for even skipped regulation in the entire *Drosophila* embryo. eLife 2013;2. http://dx.doi.org/10.7554/eLife.00522 [n/a].

[7] Cox III RS, Surette MG, Elowitz MB. Programming gene expression with combinatorial promoters. Mol Syst Biol 2007;3(1). http://dx.doi.org/10.1038/msb4100187 [n/a].

[8] Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 2009;457:215–8. http://dx.doi.org/10.1038/nature07521.

[9] Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol 2012;30(6):521–30. http://dx.doi.org/10.1038/nbt.2205.

[10] Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet 2013:1021–8. http://dx.doi.org/10.1038/ng.2713.

[11] Jaeger J, Manu J Reinitz. *Drosophila* blastoderm patterning. Curr Opin Genet Dev 2012;22(6):533–41. http://dx.doi.org/10.1016/j.gde.2012.10.005.

[12] Hammar P, Leroy P, Mahmutovic A, Marklund EG, Berg OG, Elf J. The lac repressor displays facilitated diffusion in living cells. Science 2012;336(6088):1595–8. http://dx.doi.org/10.1126/science.1221648.

[13] Mirny LA. Nucleosome-mediated cooperativity between transcription factors. PNAS 2010;107(52):22534–9. http://dx.doi.org/10.1073/pnas.0913805107.

[14] Wasson T, Hartemink AJ. An ensemble model of competitive multi-factor binding of the genome. Genome Res 2009;19:2101–12. http://dx.doi.org/10.1101/gr.093450.109.

[15] Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, et al. Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. PLoS Genet 2013;9(8):e1003571. http://dx.doi.org/10.1371/journal.pgen.1003571.

[16] Giniger E, Ptashne M. Cooperative DNA binding of the yeast transcriptional activator GAL4. PNAS 1988;85(2):382–6 [URL http://www.pnas.org/content/85/2/382.abstract].

[17] Hertel KJ, Lynch KW, Maniatis T. Common themes in the function of transcription and splicing enhancers. Curr Opin Cell Biol 1997;9(3):350–7. http://dx.doi.org/10.1016/S0955-0674(97)80007-5.

[18] Anderson GM, Freytag SO. Synergistic activation of a human promoter in vivo by transcription factor Sp1. Mol Cell Biol 1991;11(4):1935–43. http://dx.doi.org/10.1128/MCB.11.4.1935.

[19] He X, Samee MAH, Blatti C, Sinha S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol 2010;6(9):e1000935. http://dx.doi.org/10.1371/journal.pcbi.1000935.

[20] Lin YS, Carey M, Ptashne M, Green MR. How different eukaryotic transcriptional activators can cooperate promiscuously. Nature 1990;345(6273):359–61. http://dx.doi.org/10.1038/345359a0.

[21] Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. Homotypic regulatory clusters in *Drosophila*. Genome Res 2003;13:579–88. http://dx.doi.org/10.1101/gr.668403.

[22] Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res 2010;20(5):565–77. http://dx.doi.org/10.1101/gr.104471.109.

[23] Sinha S, Adler AS, Field Y, Chang HY, Segal E. Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res 2008;18(3):477–88. http://dx.doi.org/10.1101/gr.6828808.

[24] Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, Mandel G, et al. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res 2006;34(8):2238–46. http://dx.doi.org/10.1093/nar/gkl248.

[25] Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 2012; 30(3):265–70. http://dx.doi.org/10.1038/nbt.2136.

[26] Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol 2012;30(3):271–7. http://dx.doi.org/10.1038/nbt.2137.

[27] White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. PNAS 2013;110(29):11952–7. http://dx.doi.org/10.1073/pnas.1307449110.

[28] Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz Rascado L, Solano-Lira H, et al. RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor units). Nucleic Acids Res 2011;39(Suppl. 1):D98–D105.

[29] Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. Nucleic Acids Res 2011;39(Suppl. 1):D118–23.

[30] Chu D, Zabet NR, Mitavskiy B. Models of transcription factor binding: sensitivity of activation functions to model assumptions. J Theor Biol 2009;257(3):419–29. http://dx.doi.org/10.1016/j.jtbi.2008.11.026.

[31] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: models. Curr Opin Genet Dev 2005;15:116–24. http://dx.doi.org/10.1016/j.gde.2005.02.007.

[32] Zabet NR, Chu DF. Computational limits to binary genes. J R Soc Interface 2010;7: 945–54. http://dx.doi.org/10.1098/rsif.2009.0474.

[33] Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. PNAS 2003;100(9):5136–41. http://dx.doi.org/10.1073/pnas.0930314100.

[34] Mayo AE, Setty Y, Shavit S, Zaslaver A, Alon U. Plasticity of the cis-regulatory input function of a gene. PLoS Biol 2006;4(4):e45. http://dx.doi.org/10.1371/journal.pbio.0040045.

[35] Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al. Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. Mol Cell 2010;37(3): 418–28. http://dx.doi.org/10.1016/j.molcel.2010.01.016.

[36] Atkinson TJ, Halfon MS. Regulation of gene expression in the genomic context. Comput Struct Biotechnol J 2014;9(13):1–21. http://dx.doi.org/10.5936/csbj.201401001.

[37] Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. PNAS 2012;109(50):20461–6. http://dx.doi.org/10.1073/pnas.1209312109.

[38] Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by lambda phage repressor. PNAS 1982;79:1129–33 [URL http://www.pnas.org/content/79/4/1129.abstract].

[39] Whitington T, Frith MC, Johnson J, Bailey TL. Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res 2011;39(15):e98. http://dx.doi.org/10.1093/nar/gkr341.

[40] Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, et al. The role of binding site cluster strength in bicoid-dependent patterning in *Drosophila*. PNAS 2005;102(14):4960–5. http://dx.doi.org/10.1073/pnas.0500373102.

[41] Vashee S, Melcher K, Ding W, Johnston SA, Kodadek T. Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein–protein interactions. Curr Biol 1998;8(8):452–8. http://dx.doi.org/10.1016/S0960-9822(98)70179-4.

[42] Segal E, Widom J. From DNA sequence to transcriptional behaviour: a quantitative approach. Nat Rev Genet 2009;10:443–56. http://dx.doi.org/10.1038/nrg2591.

[43] Riggs AD, Bourgeois S, Cohn M. The lac represser–operator interaction: III. Kinetic studies. J Mol Biol 1970;53(3):401–17. http://dx.doi.org/10.1016/0022-2836(70)90074-4.

[44] Zabet NR, Adryan B. Computational models for large-scale simulations of facilitated diffusion. Mol BioSyst 2012;8(11):2815–27. http://dx.doi.org/10.1039/C2MB25201E.

[45] Berg OG, Winter RB, von Hippel PH. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry 1981;20(24):6929–48.

[46] von Hippel PH, Berg OG. Facilitated target location in biological systems. J Biol Chem 1989;264(2):675–8 [URL http://www.jbc.org/content/264/2/675.abstract].

[47] Halford SE. An end to 40 years of mistakes in DNA–protein association kinetics? Biochem Soc Trans 2009;37:343–8. http://dx.doi.org/10.1042/BST0370343.

[48] Kim JG, Takeda Y, Matthews BW, Anderson WF. Kinetic studies on Cro repressor–operator DNA interaction. J Mol Biol 1987;196(1):149–58. http://dx.doi.org/10.1016/0022-2836(87)90517-1.

[49] Shimamoto N. One-dimensional diffusion of proteins along DNA. J Biol Chem 1999; 274(22):15293–6. http://dx.doi.org/10.1074/jbc.274.22.15293.

[50] Kabata OKH, Arai MWI, Margarson SA, Glass RE, Shimamoto N. Visualization of single molecules of RNA polymerase sliding along DNA. Science 1993;262(5139): 1561–3. http://dx.doi.org/10.1126/science.8248804.

[51] Blainey PC, van Oijen AM, Banerjee A, Verdine GL, Xie XS. A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. PNAS 2006;103(15):5752–7. http://dx.doi.org/10.1073/pnas.0509723103.

[52] Leith JS, Tafvizi A, Huang F, Uspal WE, Doyle PS, Fersht AR, et al. Sequence-dependent sliding kinetics of p53. PNAS 2012;109(41):16552–7. http://dx.doi.org/10.1073/pnas.1120452109.

[53] Elf J, Li G-W, Xie XS. Probing transcription factor dynamics at the single-molecule level in a living cell. Science 2007;316:1191–4. http://dx.doi.org/10.1126/science.114196.

[54] Vukojevic V, Papadopoulos DK, Terenius L, Gehring WJ, Rigler R. Quantitative study of synthetic Hox transcription factor–DNA interactions in live cells. PNAS 2010; 107(9):4093–8. http://dx.doi.org/10.1073/pnas.0914612107.

[55] Chen J, Zhang Z, Li L, Chen B-C, Revyakin A, Hajj B, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell 2014;156(6):1274–85. http://dx.doi.org/10.1016/j.cell.2014.01.062.

[56] Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science 2011;332(6028):475–8. http://dx.doi.org/10.1126/science.1202142.

[57] Wang F, Redding S, Finkelstein IJ, Gorman J, Reichman DR, Greene EC. The promoter-search mechanism of *Escherichia coli* RNA polymerase is dominated by three-dimensional diffusion. Nat Struct Mol Biol 2013;20(2):174–81. http://dx.doi.org/10.1038/nsmb.2472.

[58] Hoffman MM, Birney E. An effective model for natural selection in promoters. Genome Res 2010;20(5):685–92. http://dx.doi.org/10.1101/gr.096719.109.

[59] Ruusala T, Crothers DM. Sliding and intermolecular transfer of the lac repressor: kinetic perturbation of a reaction intermediate by a distant DNA sequence. PNAS 1992;89(11):4903–7 [URL http://www.pnas.org/content/89/11/4903.abstract].

[60] Sharon E, van Dijk D, Kalma Y, Keren L, Yakhini OMZ, Segal E. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res 2014. http://dx.doi.org/10.1101/gr.168773.113 [n/a].

[61] Brackley CA, Cates ME, Marenduzzo D. Facilitated diffusion on mobile DNA: configurational traps and sequence heterogeneity. Phys Rev Lett 2012;109(16):168103. http://dx.doi.org/10.1103/PhysRevLett.109.168103.

[62] Weindl J, Dawy Z, Hanus P, Zech J, Mueller JC. Modeling promoter search by *E. coli* RNA polymerase: one-dimensional diffusion in a sequence-dependent energy landscape. J Theor Biol 2009;259(3):628–34.

[63] Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J, Kosmrlj A. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. J Phys A Math Theor 2009;42:434013. http://dx.doi.org/10.1088/1751-8113/42/43/434013.

[64] Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science (New York, NY) 2007;315(5809):233–7. http://dx.doi.org/10.1126/science.1131007.

[65] Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, et al. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. PNAS 2012;109(52):21330–5. http://dx.doi.org/10.1073/pnas.1209589110.

[66] Hermsen R, Tans S, ten Wolde PR. Transcriptional regulation by competing transcription factor modules. PLoS Comput Biol 2006;2:1552–60. http://dx.doi.org/10.1371/journal.pcbi.0020164.

[67] Lusk RW, Eisen MB. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. PLoS Genet 2010;6(1):e1000829. http://dx.doi.org/10.1371/journal.pgen.1000829.

[68] He X, Duque TSPC, Sinha S. Evolutionary origins of transcription factor binding site clusters. Mol Biol Evol 2012;29(3):1059–70. http://dx.doi.org/10.1093/molbev/msr277.

[69] Smith T, Husbands P, Layzell P, O'Shea M. Fitness landscapes and evolvability. Evol Comput 2002;10(1):1–34. http://dx.doi.org/10.1162/106365602317301754.

[70] Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet 2009;25(10):434–40. http://dx.doi.org/10.1016/j.tig.2009.08.003.

[71] Polavarapu N, Mariño Ramirez L, Landsman D, McDonald JF, Jordan IK. Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics 2008;9:226.

[72] Nourmohammad A, Lässig M. Formation of regulatory modules by local sequence duplication. PLoS Comput Biol 2011;7(10):e1002167.

[73] Hammar P, Walldén M, Fange D, Persson F, Baltekin O, Ullman G, et al. Direct measurement of transcription factor association excludes a simple operator occupancy model for gene regulation. Nat Genet 2014;46:405–8. http://dx.doi.org/10.1038/ng.2905.

[74] Mueller F, Wach P, McNally JG. Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. Biophys J 2008;94(8):3323–39. http://dx.doi.org/10.1529/biophysj.107.123182.

[75] Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature 2012;482(7385):331–8. http://dx.doi.org/10.1038/nature10886.