Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.09.002.

References
1 Miller, W. *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456, 387–390
2 Krause, J. *et al.* (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439, 724–727
3 Rogaev, E.I. *et al.* (2006) Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biol.* 4, e73
4 Green, R.E. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444, 330–336
5 Noonan, J.P. *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* 314, 1113–1118
6 Green, R.E. *et al.* (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134, 416–426
7 Krings, M. *et al.* (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* 90, 19–30
8 Stern, D.L. and Orgogozo, V. (2008) The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177
9 Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36
10 Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8, 206–216
11 Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell* 117, 185–198
12 Gu, X. (2004) Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* 167, 531–542
13 Khaitovich, P. *et al.* (2004) A neutral model of transcriptome evolution. *PLoS Biol* 2, E132
14 Liao, B.Y. and Zhang, J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* 23, 530–540
15 Gu, Z. *et al.* (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18, 609–613
16 Kaessmann, H. *et al.* (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10, 19–31
17 Shemesh, R. *et al.* (2006) Genomic fossils as a snapshot of the human transcriptome. *Proc. Natl Acad. Sci. USA* 103, 1364–1369
18 Graur, D. *et al.* (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* 28, 279–285
19 Pavlicek, A. *et al.* (2006) Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends Genet.* 22, 69–73
20 Zhang, Z. *et al.* (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* 20, 62–67
21 Balasubramanian, S. *et al.* (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.* 10, R2
22 Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120
23 Ronfani, L. *et al.* (2001) Reduced fertility and spermatogenesis defects in mice lacking chromosomal protein Hmgb2. *Development* 128, 1265–1273
24 Sakharkar, K.R. *et al.* (2006) Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Lett.* 580, 1472–1478
25 Zhang, J. and Webb, D.M. (2003) Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc. Natl Acad. Sci. USA* 100, 8337–8341
26 Li, W-H. (1997) *Molecular Evolution*, Sinauer, p. 184
27 Schrimpf, S.P. *et al.* (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 7, e48

Different gene regulation strategies revealed by analysis of binding motifs

Zeba Wunderlich[1] and Leonid A. Mirny[2]

[1] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2] Harvard–MIT Division of Health Sciences and Technology, Massachusetts, Institute of Technology, Cambridge, MA 02139, USA

Coordinated regulation of gene expression relies on transcription factors (TFs) binding to specific DNA sites. Our large-scale information–theoretical analysis of >950 TF-binding motifs demonstrates that prokaryotes and eukaryotes use strikingly different strategies to target TFs to specific genome locations. Although bacterial TFs can recognize a specific DNA site in the genomic background, eukaryotic TFs exhibit widespread, non-functional binding and require clustering of sites to achieve specificity. We find support for this mechanism in a range of experimental studies and in our evolutionary analysis of DNA-binding domains. Our systematic characterization of binding motifs provides a quantitative assessment of the differences in transcription regulation in prokaryotes and eukaryotes.

DNA binding and gene regulation
Classical experiments have demonstrated that strong binding of a TF to its cognate site in a promoter is sufficient to alter gene expression [1]. Significant effort has been put into experimentally determining [2–6] and computationally inferring [7–10] motifs recognized by TFs, and determining the occupancy of promoters by TFs [11]. The motifs and binding locations of a TF have in turn been used to predict which genes it regulates and their expression levels [12]. Such studies rely on linking the binding of TFs to DNA with the regulation of nearby genes.

Corresponding author: Mirny, L.A. (leonid@mit.edu).

Although such an association has been strongly established in bacteria, growing experimental evidence in eukaryotes challenges this assumption by showing a limited correlation between gene expression and TF binding [12–14]. For example, Gao *et al.* found no correlation between occupancy patterns and gene expression profiles for the majority (67%) of yeast TFs they studied, suggesting that only a subset of promoters bound by each TF is controlled by it [12]. A more striking example comes from a recent study [13], which demonstrated only a 3% overlap between TF occupancy and gene response to TF knockout. Although this discrepancy can be partially explained by the redundant binding of homologous TFs [15], it might also be evidence of a more fundamental uncoupling between TF binding and gene expression in eukaryotes.

Our analysis of 969 TF-binding motifs provides strong support for the uncoupling hypothesis by demonstrating that eukaryotic TFs do not recognize DNA with sufficient specificity (i.e. do not possess sufficient information) to bind to cognate sites exclusively; instead they occupy tens of thousands of decoy sites throughout a genome. Although managing such promiscuous binding requires several costly mechanisms, its advantages for eukaryotes are yet to be understood.

An information–theoretical approach to binding-motif recognition

To bind its cognate site, a TF has to recognize it among $\sim 10^6$ alternative sites in bacteria or $\sim 10^9$ sites in eukaryotes.

Using information theory, we ask whether individual TFs possess enough information for such remarkably precise recognition. The application of information theory to protein–DNA recognition has a rich history [16–18] and provides a theoretical basis for current efforts to characterize motifs recognized by DNA-binding proteins using a range of *in vivo* and *in vitro* techniques [6]. The most common use of information theory is to construct 'sequence logos' that demonstrate the relative contribution of individual base pair positions to binding specificity (Figure 1). Information theory, also allows us to test whether the total information contained in a motif is sufficient to guide a protein to a specific place in a large genome.

Information theory dictates that finding a unique object among N alternatives requires $I_{\min} = \log_2 N$ bits of information (Figure 1) [19]. Similarly, a minimum of $I_{\min} = \log_2 N$ bits of information is needed to specify a unique address in a genome containing N possible sites for a TF to bind (i.e. N bp). For bacteria, with $N = 10^6$–10^7 bp this yields $I_{min} = 20$–23 bits ($I_{min} = 22$ bits for *Escherichia coli*). For eukaryotic genomes, $N = 10^8$–10^{10} bp leads to $I_{min} \approx 27$–33 bits ($I_{min} = 24$ bits for *Saccharomyces cerevisiae*, $I_{min} = 27$ bits for *Drosophila melanogaster* and $I_{min} = 31$ bits for *Homo sapiens*).

To test whether TF motifs contain enough information to identify unique sites in their corresponding genomes, we calculate the information content of 969 experimentally determined bacterial and eukaryotic motifs. As a measure of the information contained in a motif, we
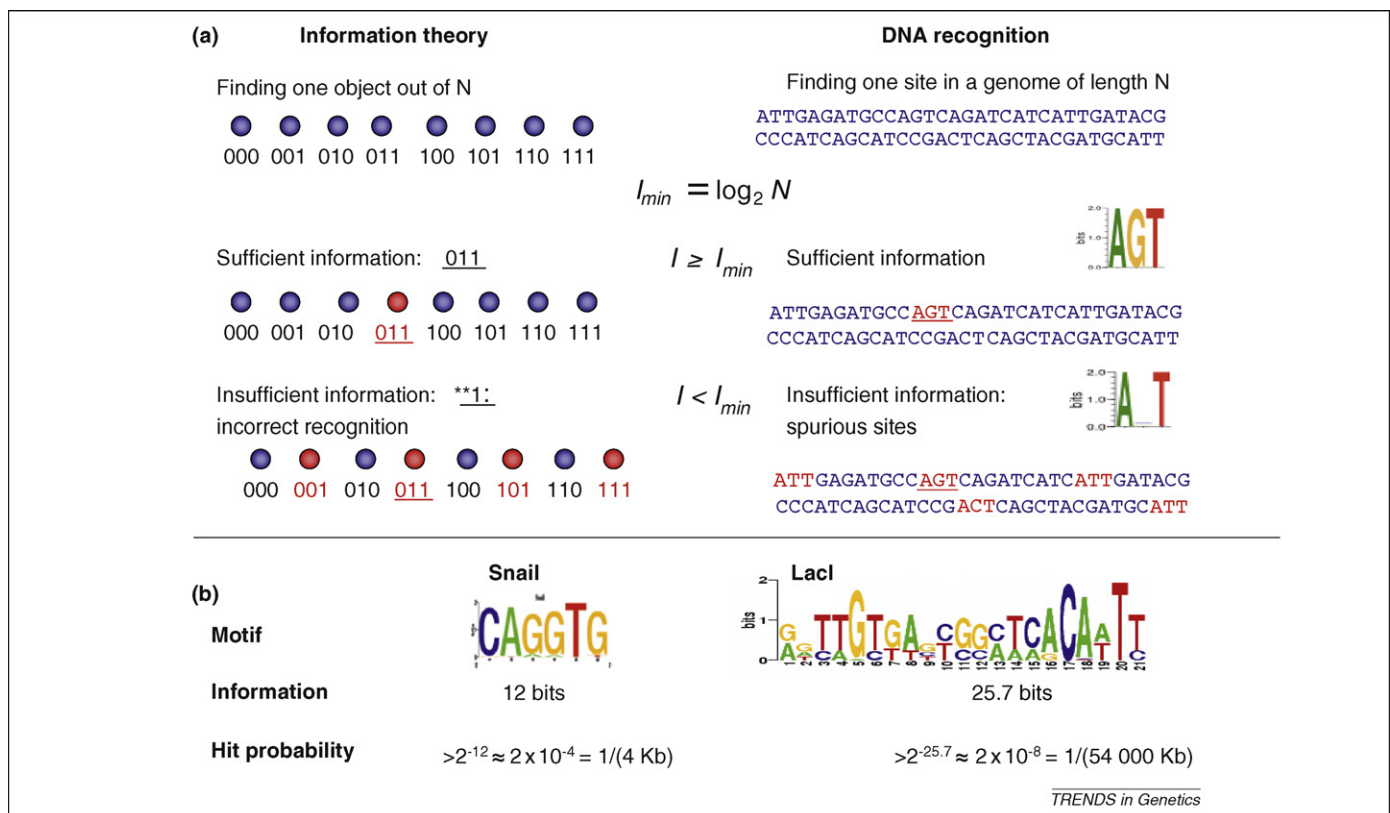

Figure 1. Information theory as applied to DNA-binding motifs. **(a)** The concepts of minimal information required in theory and in DNA recognition and the consequences of information deficiency, which results in spurious hits. **(b)** The sequence logos for low- and high-information motifs and the likelihood of a spurious hit to the motif in a 'random' genomic background.

apply the commonly used Kullback–Leibler (KL) distance between the motif and the overall genome composition [17,18]

$$I = \sum_{i=1}^{L} \sum_{b \in \{A,C,G,T\}} p_i(b) \log_2(p_i(b)/q(b)) \qquad \text{(Eq. 1)}$$

where L is the length of the motif, $p_i(b)$ is the frequency of base b at position i in the motif and $q(b)$ is its background frequency. The information content of a motif quantifies the sensitivity of TF binding affinity to variation in the binding site sequence from the consensus sequence and the probability of a site occurring in a 'random' stretch of DNA [16].

Motifs of bacterial and eukaryotic TFs are markedly different

Using this metric, we find that the motifs of prokaryotic and eukaryotic TFs are strikingly different (Figure 2, Tables S5–6 in the online supplementary material). The average information content of a prokaryotic motif $I \approx 23$ bits is slightly above the required $I_{min} = 22$ bits, demonstrating that a single cognate site is generally sufficient to address a TF to a specific location in prokaryotes, though there still might be an overlap between the background and some weak but functional sites (Figure S1 in the online supplementary material).

Although longer eukaryotic genomes require a TF to be more specific, we find that eukaryotic TFs are much less specific than bacterial TFs and do not contain sufficient information to find a cognate site among 10^9 decoys. The average information content of a multicellular eukaryotic motif is only $I \approx 12.1$ bits, falling far below the $I_{min} \approx 30$ bits required to provide a specific address in a eukaryotic genome (Figure 2). Yeast TF motifs have a mean information content of $I = 13.8$ bits, which is below the required $I_{min} \approx 24$ bits, but represents a smaller information deficiency ($I_{min} - I \approx 10$ bits) than that of the multicellular eukaryotes ($I_{min} - I \approx 18$ bits).

To ensure that the results were not influenced by a poor choice of data, we employ databases [20,21] that contain motifs for full biological TF units (i.e. dimers when the
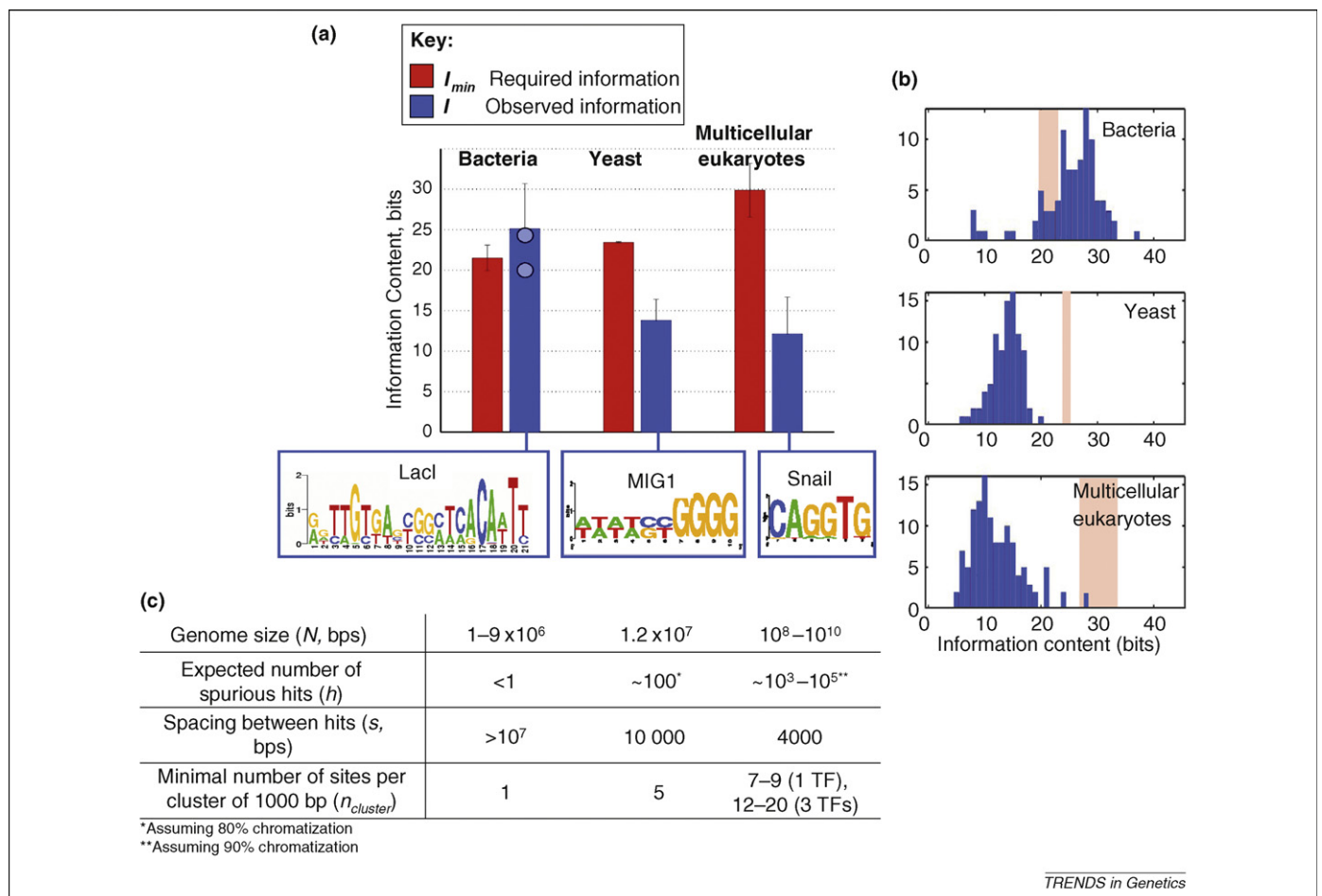

Figure 2. Properties of binding motifs for bacteria, yeast and multicellular eukaryotes. **(a)** The bar chart displays the minimum required information content for bacteria, yeast and multicellular eukaryotes (red), and the mean information content of TF-binding motifs (blue) for 98 bacterial [21], 124 yeast [22] and 123 multicellular [20] eukaryotic motifs. The error bars are ± 1 SD for the information content and for I_{min} the error bars represent the variability in that quantity because of the range of genome sizes N. The blue circles on top of the bar for bacteria indicate the average information content from several other TF-binding motif databases (Table S6). Below each series in the bar chart, we display an example of the sequence logo for a binding motif with close to average information content. The chart demonstrates that bacterial TF-binding motifs are informative enough to make spurious hits to the genomic background unlikely, in contrast to yeast and multicellular eukaryotic motifs. **(b)** The distributions of information content of motifs from the three representative databases cited above. The ranges of required information (I_{min}) are marked in red. Most bacterial motifs have $I > I_{min}$, whereas almost all eukaryotic motifs do not. **(c)** The average properties of TF-binding motifs, and the expected number and spacing between the spurious sites per genome in bacteria, yeast and multicellular eukaryotes.

binding of an individual site is accomplished by a dimer, e.g. LacI, Gal4). We also rely on *in vitro* experiments [22] that used full-length TFs. In addition, the motifs do not show a significant correlation between the information content and the number of cognate sites used to derive the motif ($\rho = -0.27$). When motifs with <8 cognate sites in RegTransBase are eliminated, we see a decrease in the mean information content by ~1 bit. Taken together, we conclude the biases from the number of sites used to construct a TF-binding motif do not change our general findings. Finally, these results are consistent for motifs obtained both *in vivo* and *in vitro* and for all available data sets (Table S6 in the supplementary material).

Widespread nonfunctional binding in multicellular eukaryotes

The significant information deficiency in eukaryotes, which emerges because of their large genomes and the degeneracy of the motifs, has several biologically important consequences. Primarily, it suggests that numerous sites as strong as the cognate ones are expected to be present in eukaryotic genomes by chance. Using information theory and simulations, we estimate the lower bound of the number of such spurious sites or hits as $h \geq 2^{I_{min}-1}$, with an average spacing $s \leq 2^I$ between them (Figure S1c in the supplementary material). Therefore, an average multicellular eukaryotic TF is expected to have $h \approx 10^4$–10^6 spurious sites per genome, which is reduced to $h \approx 10^3$–10^5 accessible sites assuming 90% chromatinization of the genome or $h \approx 10^2$–10^4 assuming 98% chromatinization. For yeast, the figure is $h \approx 10^2$–10^4, assuming 0 to 80% chromatinization.

In multicellular eukaryotes, spurious sites are expected to arise by chance every $s \approx 4$ kb. An important implication of this is that, in eukaryotes, the presence of a site cannot be a distinctive feature of a regulatory region. By contrast, a typical bacterial TF is expected to have few such spurious sites, making the presence of a single high-affinity site a unique event and a distinctive feature of a regulatory region. Consistent with this picture is the atypically low information content of a few bacterial DNA-binding proteins that pack and crosslink DNA: H-NS (histone-like nucleoid structuring protein), Fis (factor for inversion stimulation) and IHF (integration host factor) ($I = 7.5$, 7.3 and 7.8 bits, respectively). Similarly, and in agreement with Sengupta el al. [8], CRP (catabolism repressor protein) and other global regulators that bind hundreds of sites in the genome have lower information content (CRP: $I = 11$ bits). The low information content of bacterial global regulator motifs makes it particularly challenging to find their cognate sites [23].

Because information–theoretical results depend on a rather simple description of the genomic background, we searched real genomic sequences for matches to several well-characterized motifs. Using a standard bioinformatics approach, we find, in agreement with the theory, >10^4 spurious sites per genome for degenerate eukaryotic TFs (Table S1). This in no way constrains the number of cognate, functional sites a TF has in the genome but demonstrates that, in eukaryotes, cognate sites can be difficult to recognize among 10^3–10^5 equally strong spurious sites.

This creates a binding landscape with a potential for widespread nonfunctional binding.

Widespread nonfunctional binding is consistent with diverse experimental data

Evidence of this landscape has been found in several large-scale experiments. Our estimate of ~10^3 spurious hits in the chromatinized *D. melanogaster* genome is consistent with the 10^3–10^4 experimentally observed binding events for several TFs [14]. Moreover, our results explain the large number of binding events detected by ChIP-chip [11] and ChIP-seq experiments [24], suggesting that the majority of these events reflect the widespread binding to sites that arise by chance and are likely to be nonfunctional. In agreement with this idea, studies in yeast have shown a decoupling between binding and apparent regulatory function for a nontrivial fraction of TF binding events [12,13].

Using the estimated frequency of spurious sites in multicellular eukaryotes of once every 4 kb, and assuming a regulatory (accessible) region of ~1 kb around the transcription start site of each gene, we estimate that a single TF is expected to bind spuriously to ~25% of all regulatory regions. Consistent with these estimates, ChIP-chip experiments found that NOTCH1 binds to 19%, MYC to 48% and HES1 to 18% of all human promoters [25]. Our expectation is that most of these binding events have little regulatory effect. The prevalence of widespread, spurious binding events in eukaryotes means that we should be cautious in interpreting all experimentally identified binding events as regulatory interactions.

The abundance of accessible high-affinity spurious sites in eukaryotes has two effects: (i) it sequesters TF molecules; and (ii) it makes it more difficult for the cellular machinery of gene regulation to detect regulatory regions occupied by TFs and discriminate them from occupied spurious sites.

The sequestration of TF molecules by spurious binding sites necessitates a high TF copy number. The number of spurious sites h (or the number of cognate sites to be bound) imposes a lower limit on the TF copy number per cell [26], which is of the order of 1–10 per cell for bacteria, 1000 for yeast and 10^3–10^5 for multicellular eukaryotes. These estimates are consistent with available experimental data: 5–10 copies per cell of LacI repressor in *E. coli*; an average of approximately 2000 copies per cell of TFs in yeast; and 10^5 copies per cell of the prototypical multicellular eukaryotic TF p53 (Table S4).

Clustering of cognate sites can provide specificity of eukaryotic TFs

Although high TF copy numbers are necessary to cope with spurious sites, they are not sufficient to provide specificity (i.e. to allow cellular machinery to distinguish regulatory binding sites from equally strong decoys). However, the presence of multiple sites in proximity to each other can specify a regulatory region. Many regulatory regions in eukaryotes contain multiple sites of the same or different TFs [7,27–35], a property commonly used in bioinformatics to detect regulatory regions [27,31]. Using the information content of TF motifs, we can calculate the minimal number

of cognate sites ($n_{cluster}$) in regions of length $w \approx 500$–1000 bp needed to determine a unique location in a genome (supplementary methods online, Tables S2 and S3). To obtain $n_{cluster}$, we first calculate how many clusters of n spurious sites are expected to be found in a genome of a given length $E(n)$. Next, we choose $n_{cluster}$ as the minimum number of sites in a cluster such that $E(n) < 1$. In other words, a cluster of sites is unique (i.e. informative) if spurious sites are expected to form less than one such cluster by chance.

In a region of 1 kb composed of the sites of 3–10 different TFs, we calculate $n_{cluster} = 10$–20 sites. This lower limit on the number of required binding sites is remarkably consistent with the mean of ~20 sites per 1 kb observed in fly developmental enhancers [28]. These results also demonstrate that, beyond the known examples in flies and sea urchins [35], clustering of sites is a common phenomenon applicable to many regulatory regions of multicellular eukaryotes.

We also use an information–theoretical approach to calculate the information content of a cluster of sites, and then estimate the minimum number of sites in each cluster sufficient to reach the required information I_{min}. We demonstrate (see online supplementary material) that for a cluster of sites spanning a region of w bp, the

contribution of each site i to the total information content of the cluster (δI_i) is approximately

$$\delta I_i \approx I_i - \log_2 w \qquad \text{(Eq. 2)}$$

where I_i is the information content of motif i. Choosing $w = 500$–1000 bp [31,36] and $I_i = 12$ bits, we obtain that each site contributes 2–3 bits of information, necessitating 10–15 sites to achieve the ~30 bits of information needed for multicellular eukaryotes.

Eukaryotic and bacterial TFs using different repertoires of DNA-binding domains

Our study shows that combinatorial regulation is rooted in the way eukaryotic TFs recognize DNA, but how did this difference from prokaryotes arise? The gradual modifications of DNA-binding residues, expansion or contraction of the DNA-binding interface and/or re-invention of DNA-binding domains might have contributed to this difference. To investigate the possible evolutionary trajectory, we compare sequences of prokaryotic and eukaryotic DNA-binding domains of TFs available in the PFAM database [37] (Figure 3a). This analysis gives a clear result – prokaryotes and eukaryotes use different sets of DNA-binding domains. Of the 133 known DNA-binding domains, 69 have only eukaryotic members, 49 are totally prokaryotic and
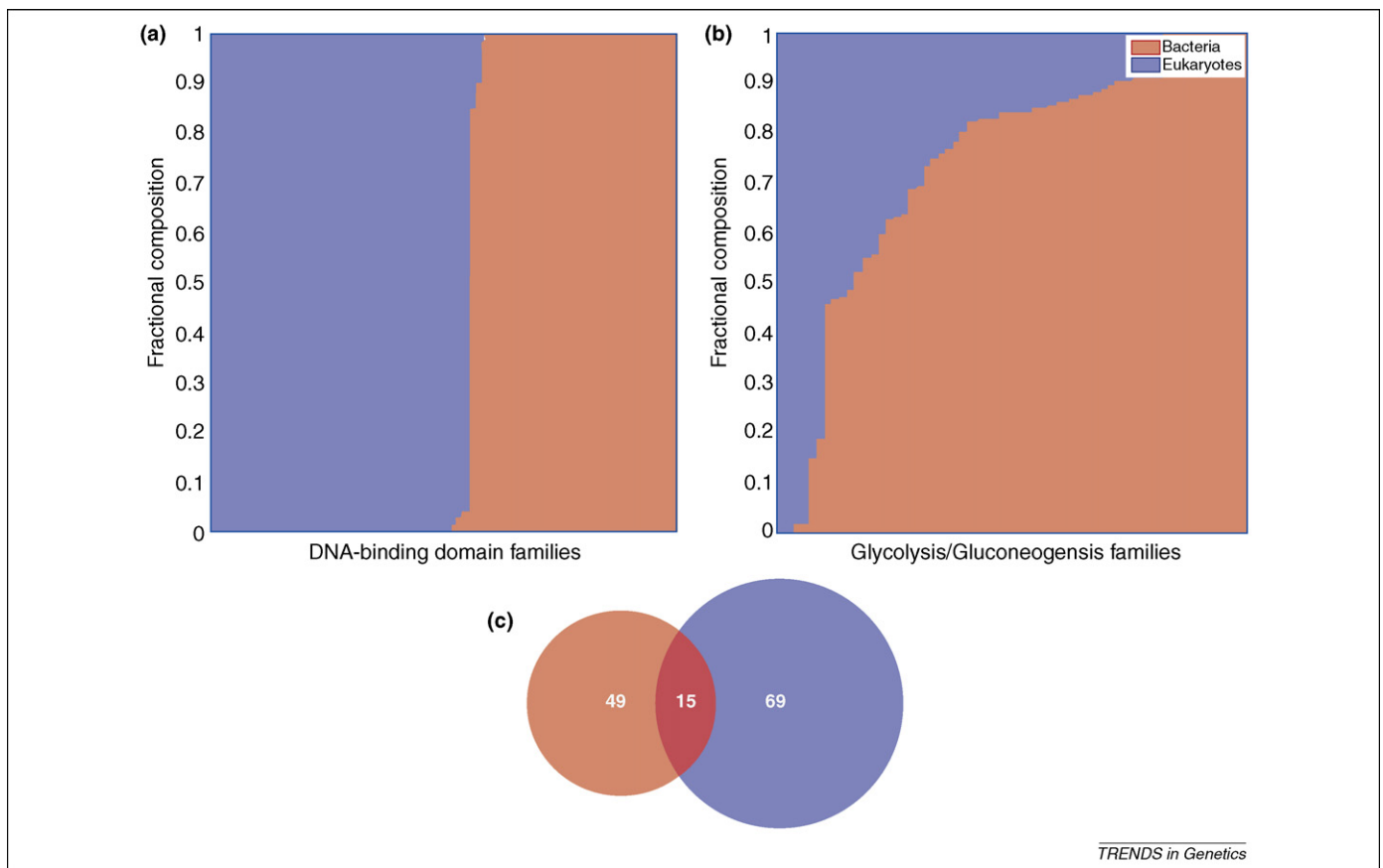

Figure 3. Membership of PFAM protein domain families by kingdom. To explore the evolution of DNA-binding domains, we examined the membership of PFAM protein domain families. Each column in **(a,b)** represents a single PFAM family, and the size of the red or blue bar indicates the proportion of the family's bacterial and eukaryotic members, respectively. **(a)** shows the membership of DNA-binding domains, demonstrating that bacteria and eukaryotes share very few. As a control **(b)**, we plot the composition of PFAM glycolysis and/or gluconeogenesis enzyme families, which are shared between kingdoms. In **(c)**, we show a Venn diagram after removing the weakest 10% of hits to a PFAM family profile.

only 15 families have both prokaryotic and eukaryotic members, but are usually dominated by one of two kingdoms (Table S7). This result is consistent with the previous observation of the differing rates of expansion and contraction of DNA-binding domain families between prokaryotes and eukaryotes [38]. As a control, we compare this result to domains involved in glycolysis and gluconeogenesis and find that a few of those domains are kingdom specific (Figure 3b). The lack of shared prokaryotic and eukaryotic DNA-binding domain families suggests that the TF machinery employed by eukaryotes might have evolved *de novo*.

Energy-based considerations of TF binding

As demonstrated in the seminal paper by Berg and von Hippel [16] and later papers this information–theoretical approach is closely related to the energy-based analysis of TF-binding motifs. The constraints on the information content of motifs considered here can be interpreted as constraints on the sequence-specific protein–DNA binding energy. Gerland *et al.* [26] and Lassig [39] have considered these constraints and demonstrated that the energy contribution of each consensus base pair to the sequence-specific binding energy in bacteria should be approximately $\varepsilon \approx 2\text{--}3$ k_BT for a motif of $L = 15$ bp.

The specificity of TF binding can be assessed using an energy-based approach. Given a set of cognate sites, how many sites in a genome are expected to have the energy lower than the energy of the cognate sites? A direct answer is provided by our bioinformatics analysis, where such sites are explicitly counted in each genome. We also use the information content of TF motifs to estimate the contribution of each consensus base pair to the sequence-specific binding energy (supplementary methods online), obtaining a range $\varepsilon \approx 1.5\text{--}3.5$ $k_BT = 1\text{--}2$ Kcal/mol for both prokaryotes and eukaryotes, which is consistent with recent micro-fluidic measurements [2].

Another important aspect of TF recognition not considered here is the nonspecific binding of proteins to DNA because our focus was on specific (high-affinity) binding. As demonstrated previously [26,39,40], competition between specific binding to cognate sites and nonspecific binding to the rest of the DNA determines whether a TF is bound to the cognate site or to nonspecific DNA. Using available dissociation constants for specific and nonspecific binding [2,41,42], we calculate that a bacterial TF binds nonspecifically once every 10^6 bp. Eukaryotic TFs, by contrast, bind nonspecifically every $10^3\text{--}10^4$ bp. Therefore, nonspecific binding sequesters almost as many TF molecules as the spurious sites, making it difficult for the cell to recognize a regulatory region from the rest of the DNA where TFs are bound specifically and nonspecifically.

Concluding remarks

We asked whether individual TF-binding motifs possess enough information to find a cognate site in the genome. The promiscuity of eukaryotic TFs leads to widespread, likely nonfunctional, binding to decoy sites. If supported by direct experimental evidence, this conclusion will challenge our understanding of gene regulation, which was gained largely from experiments in bacterial systems and can be summarized as: one site – one TF – one binding event. In multicellular eukaryotes, this paradigm turns into: multiple sites – thousands of copies of each TF – multiple cooperative binding events. This makes one binding event necessary, but certainly not sufficient to regulate gene expression.

Such a mechanism is consistent with the concept of combinatorial gene regulation in eukaryotes, but goes further by suggesting that not only are several sites required to form a regulatory region, but also binding to individual sites is likely to be widespread and possibly nonfunctional. Cooperative binding [1] and synergetic activation [43] are likely to be some of the mechanisms employed by the cell to differentiate between individual sites and clusters.

Although the apparent paradox of information deficiency in eukaryotes can be resolved by using regulatory regions containing clusters of sites, each TF must nevertheless be present in very high copy numbers. Clearly, maintaining the tens of thousands of copies of each TF per cell needed to saturate decoy sites comes at a metabolic cost that is likely to be outweighed by the advantages of promiscuous binding that are yet to be discovered.

Evolutionary analysis supports our information–theoretical results and shows that the observed differences in DNA recognition are not specific to a few cases but are likely to span across kingdoms and constitute fundamentally different strategies of transcriptional regulation in prokaryotes and eukaryotes. The promiscuity of eukaryotic TFs is likely to constitute one of many eukaryotic evolutionary novelties, which might enable more evolvable gene regulation and, thus, be essential for the evolution of a variety of structures [44].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.08.003.

References

1 Gann, A. and Ptashne, M. (2002) *Genes & Signals*, Cold Spring Harbor Laboratory Press
2 Maerkl, S.J. and Quake, S.R. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, N.Y.)* 315, 233–237
3 Fields, D.S. *et al.* (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol.* 271, 178–194
4 Noyes, M.B. *et al.* (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* 133, 1277–1289
5 Badis, G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science (New York, N.Y.)* 324, 1720–1723
6 Bulyk, M.L. (2006) DNA microarray technologies for measuring protein–DNA interactions. *Curr. Opin. Biot.* 17, 422–430

7 Siggia, E.D. (2005) Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.* 15, 214–221

8 Sengupta, A.M. *et al.* (2002) Specificity and robustness in transcription control networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 2072–2077

9 MacIsaac, K.D. *et al.* (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7, 113

10 Kinney, J.B. *et al.* (2007) Precise physical models of protein–DNA interaction from high-throughput data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 501–506

11 Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104

12 Gao, F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* 5, 31

13 Hu, Z. *et al.* (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* 39, 683–687

14 Li, X.Y. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.* 6, e27

15 Gitter, A. *et al.* (2009) Backup in gene regulatory networks explains differences between binding and knockout results. *Mol. Syst. Biol.* 5, 276

16 Berg, O.G. and von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193, 723–750

17 Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.* 109–113

18 Schneider, T.D. *et al.* (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 415–431

19 Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, Wiley-Interscience

20 Vlieghe, D. *et al.* (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* 34, D95–97

21 Kazakov, A.E. *et al.* (2007) RegTransBase – a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.* 35, D407–412

22 Zhu, C. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 19, 556–566

23 Mustonen, V. and Lassig, M. (2005) Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15936–15941

24 Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science (New York, N.Y.)* 316, 1497–1502

25 Margolin, A.A. *et al.* (2009) ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. *Proc. Natl. Acad. Sci. U. S. A.* 106, 244–249

26 Gerland, U. *et al.* (2002) Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12015–12020

27 Rajewsky, N. *et al.* (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3, 30

28 Berman, B.P. *et al.* (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* 5, R61

29 Ochoa-Espinosa, A. *et al.* (2005) The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4960–4965

30 Emberly, E. *et al.* (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* 4, 57

31 Berman, B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U S A* 99, 757–762

32 Markstein, M. *et al.* (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc. Natl. Acad. Sci. U. S. A.* 99, 763–768

33 Sinha, S. *et al.* (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics* 5, 129

34 Hallikas, O. *et al.* (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124, 47–59

35 Davidson, E.H. (1999) A view from the genome: spatial control of transcription in sea urchin development. *Curr. Opin. Genet. Dev.* 9, 530–541

36 Koudritsky, M. and Domany, E. (2008) Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.* 36, 6795–6805

37 Finn, R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.* 36, D281–288

38 Wilson, D. *et al.* (2007) DBD taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*

39 Lassig, M. (2007) From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* 8 Suppl 6, S7

40 Kolesov, G. *et al.* (2007) How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 13948–13953

41 Revzin, A. (1990) *The Biology of Nonspecific DNA Protein Interactions*, CRC Press

42 Ozbudak, E.M. *et al.* (2004) Multistability in the lactose utilization network of *Escherichia coli*. *Nature* 427, 737–740

43 Carey, M. *et al.* (1990) A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. *Nature* 345, 361–364

44 Carroll, S.B. (2005) Evolution at two levels: on genes and form. *PLoS Biol.* 3, e245