CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Nature Precedings

Fundamentally different strategies for transcriptional regulation are revealed by analysis of binding motifs

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

[1]Biophysics Program, Harvard University, Cambridge, MA 02138
[2]Harvard-MIT Division of Health Sciences and Technology,
Massachusetts Institute of Technology, Cambridge, MA 02139
* Current Address: Department of Systems Biology, Harvard Medical School, Boston, MA 02155

To regulate a particular gene, a transcription factor (TF) needs to bind a specific genome location. How is this genome address specified amid the presence of ~10^6-10^9 decoy sites? Our analysis of 319 known TF binding motifs clearly demonstrates that prokaryotes and eukaryotes use strikingly different strategies to target TFs to specific genome locations; eukaryotic TFs exhibit widespread nonfunctional binding and require clustering of sites in regulatory regions for specificity.

The binding of a TF to its cognate site is believed to be necessary and sufficient to trigger a cascade of events that regulate gene expression. This has been proved definitively in bacteria, and it has been assumed to be true in eukaryotes. However, growing experimental data on the *in vivo* binding of eukaryotic TFs demonstrates widespread nonfunctional binding of TFs[1] and the lack of correlation between TF binding and differential gene expression[2], suggesting that TF binding and its functional consequences are uncoupled in eukaryotes. Here we propose that this uncoupling is enabled by the fundamentally different machinery used by eukaryotic TFs to recognize their sites. Strikingly, this uncoupling between binding and gene regulation in eukaryotes is evident from the motifs of TFs. By examining a collection of TF binding motifs, we demonstrate that prokaryotes and eukaryotes use markedly different strategies to "address" a particular location in the genome. The high specificity of prokaryotic TFs targets them precisely to functional cognate sites, while the low specificity of eukaryotic TFs allows non-functional binding, but also provides an opportunity for combinatorial regulation.

We use tools of information theory[3] to characterize known motifs of bacterial and eukaryotic TFs. In a genome of size N bps, a minimum of $I_{\min} = \log_2(N)$ bits of information is needed to specify a unique address in that genome. We can compare this value to the information content (I) of actual TF binding motifs using the Kullback-Leibler (KL) distance between the motif and the overall genome composition[4]:

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

Here, 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. If $I < I_{\min}$, the motif does not have enough information to specify a unique address in a random genome of size N. Insufficient information has two consequences: (i) a large number of spurious matches of the motif are present in the genome,

creating the potential for widespread nonfunctional binding; and (ii) several sites are required to specify a regulatory region.

We calculated these quantities for 319 prokaryotic and eukaryotic TFs and found that they were markedly different (Figure 1). The average information content of a prokaryotic motif, 19.8 bits, is close to the required I_{min} = 22.2 bits, showing little information deficiency and indicating that a single cognate site is sufficient to address a TF to a specific location in prokaryotes. However, the average information content of a multicellular eukaryotic motif, 12.1 bits, falls far below I_{min} ≈ 30 bits required to provide a specific address in a eukaryotic genome. The expected number of spurious motif matches, or hits, per genome can be calculated as $h = 2^{I_{\min} - I}$, and the average spacing between the hits as $s = 2^I$ (Figure 1 and Supplementary Methods). In a eukaryotic genome, $h \approx 10^4$-10^6, and even substantial chromatization (90%) only reduces h to 10^3-10^5 spurious binding sites per genome. Note that information content of a motif does not determine or constrain the number of cognate, functional sites a TF has in the genome. However, the large number of spurious high-affinity binding sites in genomes of multicellular eukaryotes creates a binding landscape with a potential for widespread non-functional binding. These information-theoretic estimates provide a lower bound on the number of spurious binding events. To test these estimates, we searched for matches to a several well-characterized motifs in real genomic sequences, and consistent with the theory, found many spurious hits to eukaryotic binding motifs (Supplementary Methods, Supplementary Table 1).

Our analysis shows that a single cognate site is sufficient to address a TF to a specific location in prokaryotes. In contrast, a typical eukaryotic TF is expected to have specific sites arising by chance every $s \approx 4000$ bps. The phenomenon of widespread non-functional binding has been recently observed experimentally for several TFs in *D. melanogaster*[1] and *S. cerevisiae*[2]. In fact, our estimate of $\sim 10^3$ spurious hits in the chromatinized fly genome agrees well with experimentally observed 10^3-10^4 binding events. Moreover, our results help to explain the large number of binding events detected by ChIP-chip[5], suggesting that the vast majority of these event reflect the widespread *specific* (high-affinity) binding of eukaryotic TFs to sites that are not functional, but inevitable appear in the genomic background. . In agreement with our results, a recent study of yeast TFs demonstrated little overlap between genes that are differentially expressed in response to TF knockout and targets bound by this TF, suggesting there are an abundance of binding events with no detectable effect on gene expression[2]. The prevalence of widespread, unavoidable, spurious binding events in eukaryotes calls for caution in interpreting all experimentally identified binding events as regulatory interactions.

The abundance of accessible high-affinity spurious sites has two effects: (i) it sequesters TF molecules, and (ii) it makes harder for the cellular machinery of gene regulation (RNA polymerase, general transcription factors, etc.) to discriminate between functionally and non-functionally bound TFs. How then is addressing achieved in eukaryotes? We suggest that clustering of binding sites in regulatory regions combined with a sufficiently high copy-number of TFs allows eukaryotes to cope with the low information content of their TF motifs.

The sequestration of TF molecules by spurious binding sites necessitates many more TF copies per cell to occupy a few cognate sites. The number of spurious sites h imposes a lower limit on

the TF copy number per cell, which is approximately 5 copies per cell for bacteria, 2000 for yeast, and 10^3-10^5 for multicellular eukaryotes (taking into account 90% chromatization). These estimates are remarkably consistent with available experimental data: 5-10 copies per cell of *Lac* repressor in *E. coli*, an average of approximately 2000 copies per cell of TFs in yeast and 10^5 copies per cell of such prototypical multicellular eukaryotic TF as p53 (see Supplementary Table 4). Although high TF copy numbers are necessary to cope with non-functional binding, they are not sufficient to provide specificity, i.e. to allow cellular machinery to distinguish functional regulatory binding events from equally strong decoys.

Many regulatory regions in eukaryotes are known to contain multiple sites of the same or different TFs (e.g., ref [6]), suggesting that the clustering of sites can be used to predict such regions[7-9]. Our analysis allows us to calculate the minimal number of immediately adjacent sites (n) needed to specify a unique location in a genome as $n = I_{\min}/I \approx 3$ for multicellular eukaryotes and $n \approx 1$ for prokaryotic motifs. If sites are not immediately adjacent, but are located within a regulatory region of 500-1000 bps, more sites are needed to make a regulatory region stand out from the spurious binding event background. Using estimated the background frequency of hits for a single TF, we calculate the minimal number of sites of this TF per cluster ($n_{cluster}$) as 7-9 for a eukaryotic regulatory region of about 1000 bp (Supplementary Methods). Such a cluster of sites is expected to appear less than once per genome due to the spurious hits, thus allowing a cell to uniquely identify a regulatory region. If a regulatory region is composed of the sites of several different TFs, then the cluster should contain at least 12-20 binding sites in a regulatory region of 1000 bps (Supplementary Methods). This lower bound on the number of required binding sites is remarkably consistent with 20-25 sites per kilobase observed in fly developmental enhancers[8]. While fly and sea urchin enhancers are known to contain clusters of TF binding sites, our results clearly demonstrate that clustering of sites should be a common phenomenon applicable to most regulatory regions and promoters in multicellular eukaryotes. Simply put, since a single eukaryotic binding motif is unable to specify a unique address in the genome, multiple binding sites must be used in order to unambiguously specify a regulation site.

What are the advantages that low-information TF motifs provide to a eukaryotic cell? First, the required clustering of sites provides a mechanism for combinatorial gene regulation: to be recognizable by a cell, a regulatory region should contain several TFs bound in close proximity, providing an opportunity for synergistic action between TFs. Second, the short motifs of eukaryotic TFs facilitate the rapid creation of new sites and rearrangement of existing sites[10], thus enabling highly evolvable gene regulation[11] and the rapid turnover of sites in regulatory regions[8,12]. And third, combinatorial regulation obtained through site clustering allows a large number of genes to be controlled by a limited repertoire of TFs[13]. Our study shows that combinatorial regulation is rooted in the way eukaryotic TFs recognize DNA.

The observed difference in genome addressing strategy may have arisen in several different ways: gradual modifications of the DNA-binding residues of TFs, the expansion or contraction of the DNA-binding interface of TFs, the preferential use of one type of DNA-binding domain (e.g. zinc fingers) over others on a kingdom-wide scale, or the re-invention of DNA-binding domains altogether. To investigate the evolutionary trajectory of eukaryotic DNA recognition, we systematically compared sequences of prokaryotic and eukaryotic DNA-binding domains of TFs available in the PFAM database (Figure 2A).

This analysis gave a surprising result – prokaryotes and eukaryotes use different sets of DNA-binding domains. Of the 133 known DNA-binding domain families, 69 have only eukaryotic members, 49 are totally prokaryotic, and only 15 families with both prokaryotic and eukaryotic members, but are usually dominated by one of two kingdoms (Supplementary Methods, Supplementary Table 2). As a positive control, we compared this result to the domains involved in glycolysis and gluconeogenesis and found that a very small number of those domains are kingdom specific (Figure 2B). The lack of shared prokaryotic and eukaryotic DNA-binding domain families suggests that the TF machinery of low-specificity binding and largely combinatorial regulation employed by eukaryotes has evolved *de novo*.

This evolutionary analysis supports our information-theoretical results and emphasizes that the observed differences in DNA recognition are not specific to a few well-characterized TFs or organisms, but are likely to span across kingdoms and constitute fundamentally different strategies to transcriptional regulation in prokaryotes and eukaryotes.

Materials and Methods

Binding motifs for *Escherichia coli* were downloaded from RegulonDB, yeast transcription factor motifs were taken from MacIsaac, et al.[14], and the JASPAR CORE collection of eukaryotic transcription factor binding motifs was downloaded from JASPAR. Background nucleotide frequencies were calculated for those organisms with completed genome sequences and the average background frequencies were used for all others. Tables with the information content and GC content for all the transcription factors are in the Supplementary Data. The differences between prokaryotic and eukaryotic binding motifs are evident when using other motif collections (Supplementary Table 3).

Acknowledgments

We thank Shamil Sunyaev, Mikahil Gelfand, Shaun Mahoney and Alex Shpunt for insightful discussions and Michael Schnall for interpretation of the information cutoff. ZW was supported by a Howard Hughes Medical Institute Predoctoral Fellowship. LM acknowledges support of *i2b2*, NIH-supported Center for Biomedical Computing at the Brigham and Women's Hospital.

References

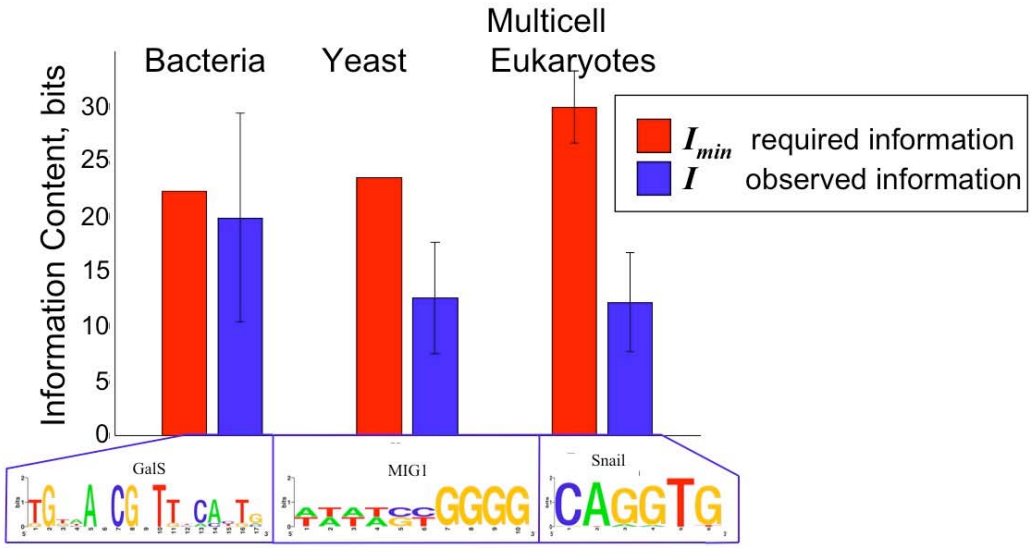
1. Li, X.Y. et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**, e27 (2008).
2. Hu, Z., Killion, P.J. & Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* **39**, 683-7 (2007).
3. Stormo, G.D. & Fields, D.S. Specificity, free energy and information content in protein-DNA interactions. in *Trends Biochem Sci* Vol. 23 109-13 (1998).
4. Schneider, T.D., Stormo, G.D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. in *J Mol Biol* Vol. 188 415-31 (1986).
5. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
6. Ochoa-Espinosa, A. et al. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc Natl Acad Sci U S A* **102**, 4960-5 (2005).

7. Emberly, E., Rajewsky, N. & Siggia, E.D. Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4**, 57 (2003).
8. Berman, B.P. et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**, R61 (2004).
9. Siggia, E.D. Computational methods for transcriptional regulation. *Curr Opin Genet Dev* **15**, 214-21 (2005).
10. Berg, J., Willmann, S. & Lassig, M. Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* **4**, 42 (2004).
11. Carroll, S.B. Evolution at two levels: on genes and form. *PLoS Biol* **3**, e245 (2005).
12. Moses, A.M. et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* **2**, e130 (2006).
13. Itzkovitz, S., Tlusty, T. & Alon, U. Coding limits on the number of transcription factors. *BMC Genomics* **7**, 239 (2006).
14. MacIsaac, K.D. et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113 (2006).
15. Wilson, D., Charoensawan, V., Kummerfeld, S.K. & Teichmann, S.A. DBD taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* (2007).

Supplementary Materials

- Supplementary Data
- Supplementary Methods
- Supplementary Table 1. Number of expected and actual hits to TF PWMs
- Supplementary Table 2. PFAM DNA-binding domain families with hits to prokaryotes and eukaryotes
- Supplementary Table 3. Average information content from other data sources
- Supplementary Table 4. The number of TF copies per cell

Fig. 1



Genome size (N)	$5 \cdot 10^6$ bps	$1.2 \cdot 10^7$ bps	$10^8 - 10^{10}$ bps
Expected number of spurious hits (h)	5	2,000	20,000 – 2,000,000
Spacing between hits (s)	9×10^5 bps	6,000 bps	4,000 bps
Minimal number of adjacent sites (n)	1	2	2-3
Minimal number of sites per cluster of 1Kbs ($n_{cluster}$)	1	4-6	7-9

Figure 1. Properties of binding sites for bacteria, yeast, and multicellular eukaryotes
The bar chart displays the minimum required information content for bacteria, yeast, and multicellular eukaryotes (red), as well as the actual information content of TF binding motifs (blue) for 72 bacterial, 124 yeast and 123 multicellular eukaryotic motifs. The error bars are ± 1 standard deviation for the actual information content, and for eukaryotic I_{min}, the error bars represent the variability in that quantity due to the range of genome sizes N. Below each series in the bar chart, we display an example of sequence logo for a binding motif with close to average information content, and other important properties of TF binding motifs.

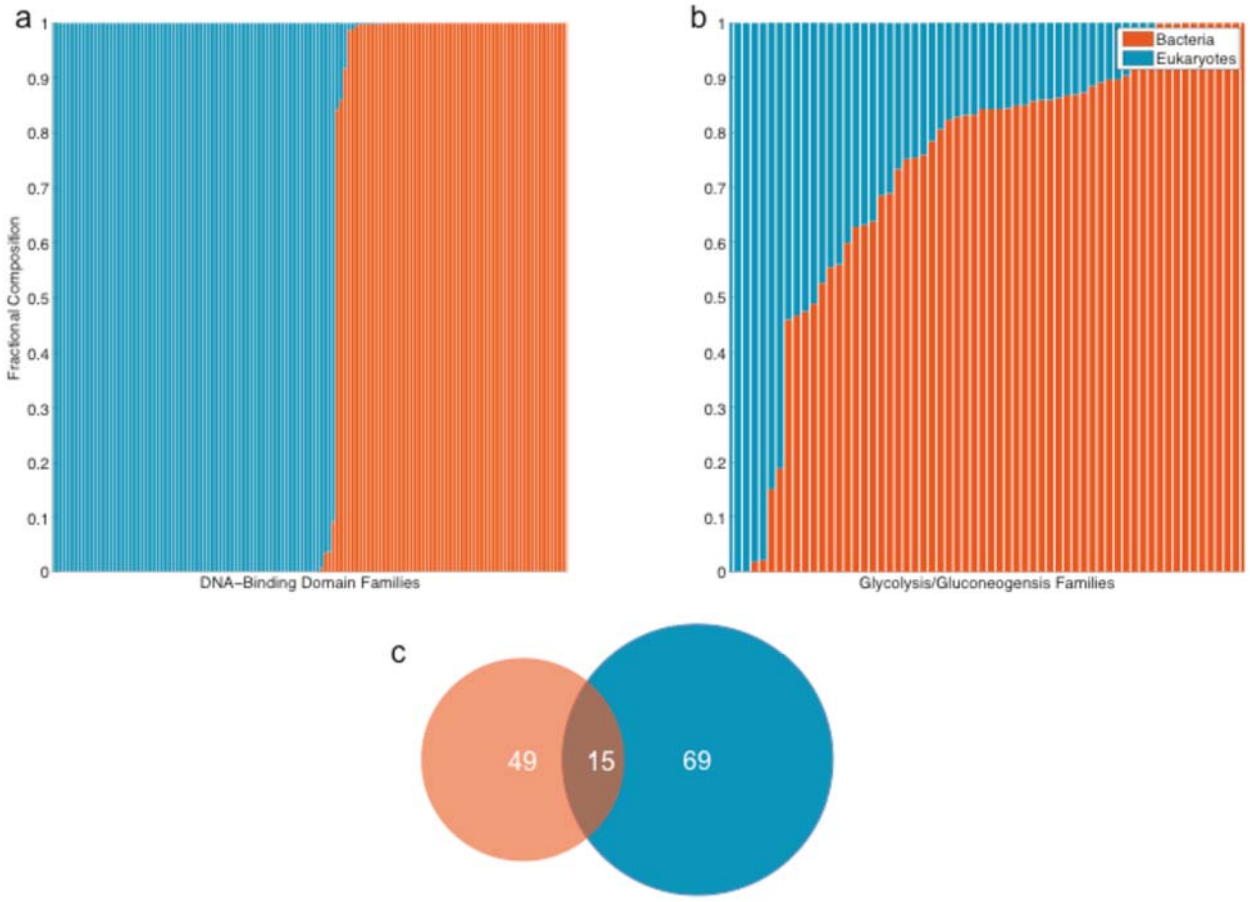


Figure 2. Membership of PFAM protein domain families, by kingdom
To explore the evolution of TF DNA-binding domains, we examined the membership of PFAM protein domain families. Each column of (a-b) represents a single PFAM family, and the size of the orange or teal bar indicates the fraction of the family's bacterial and eukaryotic members, respectively. In (a), we plot the membership of DNA-binding domains (from DBD database[15]), demonstrating that they are almost unshared by bacteria and eukaryotes, and in (c), we show a Venn diagram, after removing the weakest 10% of hits to a PFAM family profile. As a control (b), we plot the composition of PFAM glycolysis/gluconeogensis enzyme families (as reported in KEGG database), which are shared by bacteria and eukaryotes.