JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 5, 2014
© Mary Ann Liebert, Inc.
Pp. 373–384
DOI: 10.1089/cmb.2014.0032

Informational Requirements for Transcriptional Regulation

PATRICK K. O'NEILL,[1,*] ROBERT FORDER,[2,*] and IVAN ERILL[1]

ABSTRACT

Transcription factors (TFs) regulate transcription by binding to specific sites in promoter regions. Information theory provides a useful mathematical framework to analyze the binding motifs associated with TFs but imposes several assumptions that limit their applicability to specific regulatory scenarios. Explicit simulations of the co-evolution of TFs and their binding motifs allow the study of the evolution of regulatory networks with a high degree of realism. In this work we analyze the impact of differential regulatory demands on the information content of TF-binding motifs by means of evolutionary simulations. We generalize a predictive index based on information theory, and we validate its applicability to regulatory scenarios in which the TF binds significantly to the genomic background. Our results show a logarithmic dependence of the evolved information content on the occupancy of target sites and indicate that TFs may actively exploit pseudo-sites to modulate their occupancy of target sites. In regulatory networks with differentially regulated targets, we observe that information content in TF-binding motifs is dictated primarily by the fraction of total probability mass that the TF assigns to its target sites, and we provide a predictive index to estimate the amount of information associated with arbitrarily complex regulatory systems. We observe that complex regulatory patterns can exert additional demands on evolved information content, but, given a total occupancy for target sites, we do not find conclusive evidence that this effect is because of the range of required binding affinities.

Key words: binding sites, evolution, evolutionary simulation, *in silico* evolution, information content, regulation, transcription factor, transcription networks.

1. INTRODUCTION

Regulation of gene expression is a central process in living systems that is addressed primarily through the regulation of transcription (Ptashne, 2005). At its most fundamental level, transcriptional regulation is implemented by the binding of a class of proteins, known as transcription factors (TF), to specific sites in the promoter regions of regulated genes (Huang et al., 1999; Orphanides and Reinberg, 2002). Binding of a TF to these sites can either facilitate or hinder binding of the RNA–polymerase holoenzyme at the promoter, resulting in activation or repression of transcription (Ptashne, 2005; Minchin and Busby, 2009). TF-binding sites are typically short (6–20 bp) and fixed-length and often degenerate DNA sequences that are

Departments of [1]Biological Sciences and [2]Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland.

[*]These two authors contributed equally to this work and should be considered co-first authors.

efficiently located and bound by the TF through recognition of specific patterns or motifs. Information theory provides a robust theoretical framework to analyze the interaction between TF and their binding sites (Schneider et al., 1986). Given a collection of aligned TF-binding sites (i.e., a TF-binding motif), one can compute the amount of information present in the TF-binding motif as the difference in uncertainty about which base occupies each position of the motif when compared with an equivalent number of unaligned genome subsequences of the same length. More formally, the information $R_{\text{sequence}}(i)$ of the ith position in the TF-binding motif is defined as the difference between the entropies of the unaligned and aligned sequences:

$$R_{sequence}(i) = -\sum_{S}(f(S)\log_2 f(S)) - \left[-\sum_{S}(p_i(S)\log_2 p_i(S))\right]. \tag{1}$$

At any given position, the frequencies of each base in an unaligned collection of sites will follow the genomic background frequencies $f(S)$ for each base S, whereas the collection of aligned TF-binding sites will display position-specific frequencies $p_i(S)$ for the ith position. Assuming positional independence, the total amount of information embedded in the motif is

$$R_{\text{sequence}} = \sum_{i=1}^{L}(R_{\text{sequence}}(i)), \tag{2}$$

where the sum ranges over all positions i of the motif up to its length L.

The formula for R_{sequence} forms the basis of sequence logos, the standard graphical representation of TF-binding motifs, and has become the most extended model for TF-binding motifs (Schneider and Stephens, 1990). As defined, R_{sequence} can be interpreted formally as the amount of information embedded by evolution in a set of TF-binding sites, and it follows naturally that this value should be linked to the amount of information required to locate those sites in the genome. Given a genome of size G, the information required to locate M sites is given by

$$R_{\text{frequency}} = -\sum_{G}\left(\frac{1}{G}\log_2\frac{1}{G}\right) - \left[-\sum_{M}\left(\frac{1}{M}\log_2\frac{1}{M}\right)\right] = \log\frac{\mathrm{G}}{\mathrm{M}}, \tag{3}$$

which corresponds to the difference between the entropy of a system with an unspecific TF that binds with equal affinity at all genome positions and that of a system with a specific TF capable of binding with equal affinity to M target sites and nowhere else in the genome. $R_{\text{frequency}}$ therefore provides a lower bound on the information required to identify M target sites in a genome of size G.

It has been shown for several bacterial TFs operating on unbiased genomes that $R_{\text{frequency}}$ provides a good approximation to the observed R_{sequence} (Schneider et al., 1986), and the approximate equality $R_{\text{sequence}} \approx R_{\text{frequency}}$ has come to be seen as a fundamental result of the application of information theory to TFs and their binding sites. Several analytical models of TF-binding site evolution have been developed (Gerland and Hwa, 2002; Berg et al., 2004; Mustonen and Lassig, 2005; Stewart et al., 2012) and even applied to the study of the relationship between R_{sequence} and $R_{\text{frequency}}$ (Kim et al., 2003), but for tractability these models require strong assumptions for the binding model and genomic background, and thus do not capture the full dynamics of the evolution of mutual information between TFs and their cognate-binding sites. In previous work, the relationship between R_{sequence} and $R_{\text{frequency}}$ has been further substantiated by explicit evolutionary simulations (Schneider, 2000). Synthetic genomes encoding a simulated TF (or recognizer) and predefined locations for target sites were evolved using a genetic algorithm framework. The system modeled the TF as a linear classifier with an evolvable threshold capable of designating each position of the genome as bound or unbound. The fitness of an organism was evaluated as the inverse of the total amount of errors made by the recognizer: binding at nontarget locations (false positives) and failing to bind at predefined targets (false negatives). Using this model, it was shown that the observed R_{sequence} at predefined targets evolves to closely match the predicted $R_{\text{frequency}}$ (Schneider, 2000).

Most experimental techniques used traditionally in the identification of TF-binding sites provide a qualitative description of potential sites as bound or not bound by the protein. As a consequence, most theoretical models of the interaction between TFs and their binding sites have traditionally aimed at matching a binary representation of TF-binding sites. Nonetheless, as ChIP-seq and other high-throughput quantitative protein–DNA binding assays have made manifest, TFs often display a broad range of affinities for target sites and these distinct affinities lead *in vivo* to different occupancies that can have relevant physiological effects (Browning and Busby, 2004; Maerkl and Quake, 2007; Mardis, 2007). These findings have recently led to the emergence of computational methods capable of harnessing high-throughput data to capture the quantitative nature of TF activity (Roider et al., 2007; Badis et al., 2009; Zhao and Stormo, 2011). Hence, an important theoretical question in the analysis of TF-binding motifs concerns the impact of regulatory demands on the observed information content (R_{sequence}). The effect of regulatory demands on the information content of TF-binding motifs is not obvious. For instance, motifs harboring sites with lower than average occupancies might require less information because such sites need not be recognized as precisely by the TF. On the other hand, in such a situation the TF must ensure that different groups of sites be bound at distinct occupancies, which might require additional information (Erill and O'Neill, 2009). Predictive estimators of R_{sequence}, like $R_{\text{frequency}}$, are based on a binary definition of TF-binding sites and cannot be generalized trivially to systems with differential regulation of sites. In this work we develop and validate a platform to simulate the co-evolution of TFs and their binding motifs, and we use it to explore the impact of regulatory requirements in TF-binding motifs. We generalize $R_{\text{frequency}}$ to deal with realistic evolutionary scenarios, and we validate this generalization using evolutionary simulations. Our results reveal a logarithmic dependence of the evolved information content on the fraction of total probability mass associated with target sites. Further, we observe that evolved systems can effectively exploit the genomic background to enact their regulatory roles. Finally, simulations on arbitrary regulatory patterns reveal that the evolved information content is predicted by total target occupancy and is invariant with respect to the complexity of the site-wise arrangement of target occupancies.

2. RESULTS AND DISCUSSION

2.1. In silico *evolution of transcriptional regulatory networks*

The evolution of simple transcriptional regulatory networks has been studied before by means of explicit simulation of TFs and their binding sites, leading to fundamental insights into the informational constraints faced by such networks (Schneider, 2000). In this work we generalize this approach by developing an evolutionary simulator that incorporates biophysical modeling and addresses several caveats of previous attempts. Evolutionary simulator of transcriptional regulatory motifs (ESTReMo) uses a genetic algorithm backbone to simulate the evolution of transcriptional regulatory networks using a modular approach (Fig. 1). In ESTReMo, each organism contains two evolvable components: a TF model and a set of promoter regions for target genes. The TF is modeled as a feed-forward artificial neural network that operates on the set of target promoter regions and on a fixed genomic background. In each generation, the TF model is used to compute the free energy of binding for all positions in target promoter regions and in a randomly sampled

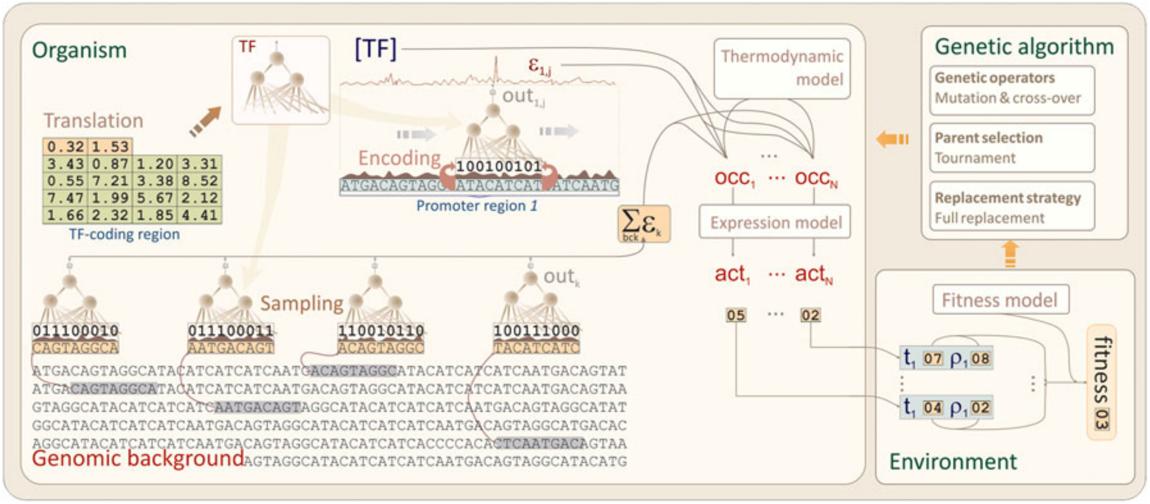

FIG. 1. Schematic representation of the evolutionary simulator of transcriptional regulatory motifs (ESTReMo). Diagram summarizes the internal organization of an individual organism and its connection with the fitness model. TF, transcription factor.

subset of the genomic background. The occupancy of each promoter position is computed according to a statistical mechanical distribution on an effective background generated by scaling up the randomly drawn samples. These occupancies are then used to determine the expression level for each target gene following a given expression model. Finally, organism fitness is evaluated as the sum of cost–benefit differences for each target gene and its desired expression level using an empirical fitness model.

The combination of thermodynamic and empirical fitness models with a modular approach taken in ESTReMo has several advantages. The interpretation of neural network outputs as binding energies permits their incorporation into a well-described biophysical framework that can explicitly model nonspecific binding and physical distributions of site occupancy, and makes the results of the simulation directly comparable with real biological systems (Stormo and Fields, 1998; Djordjevic et al., 2003). This also enables us to couple predicted expression levels with a fitness model anchored on empirical data that take into account the effects of both under- and over-regulation of target genes (Dekel and Alon, 2005). The use of continuous energy, occupancy, and expression levels in conjunction with a continuous fitness function ensures that a fitness gradient is always available for the evolutionary process. This makes it possible to evolve regulatory systems on large genomic backgrounds, which is otherwise infeasible with threshold-based approaches. Furthermore, the use of a fixed genomic background, as opposed to an evolvable one, prevents the system from actively pruning pseudo-sites and allows us to run simulations on real genome sequences.

As illustrated in Figure 2, ESTReMo simulations are characterized by a period of rapid increase in fitness (Fig. 2A), in which the TF evolves to associate strong affinities (low energy values) to a primordial core of

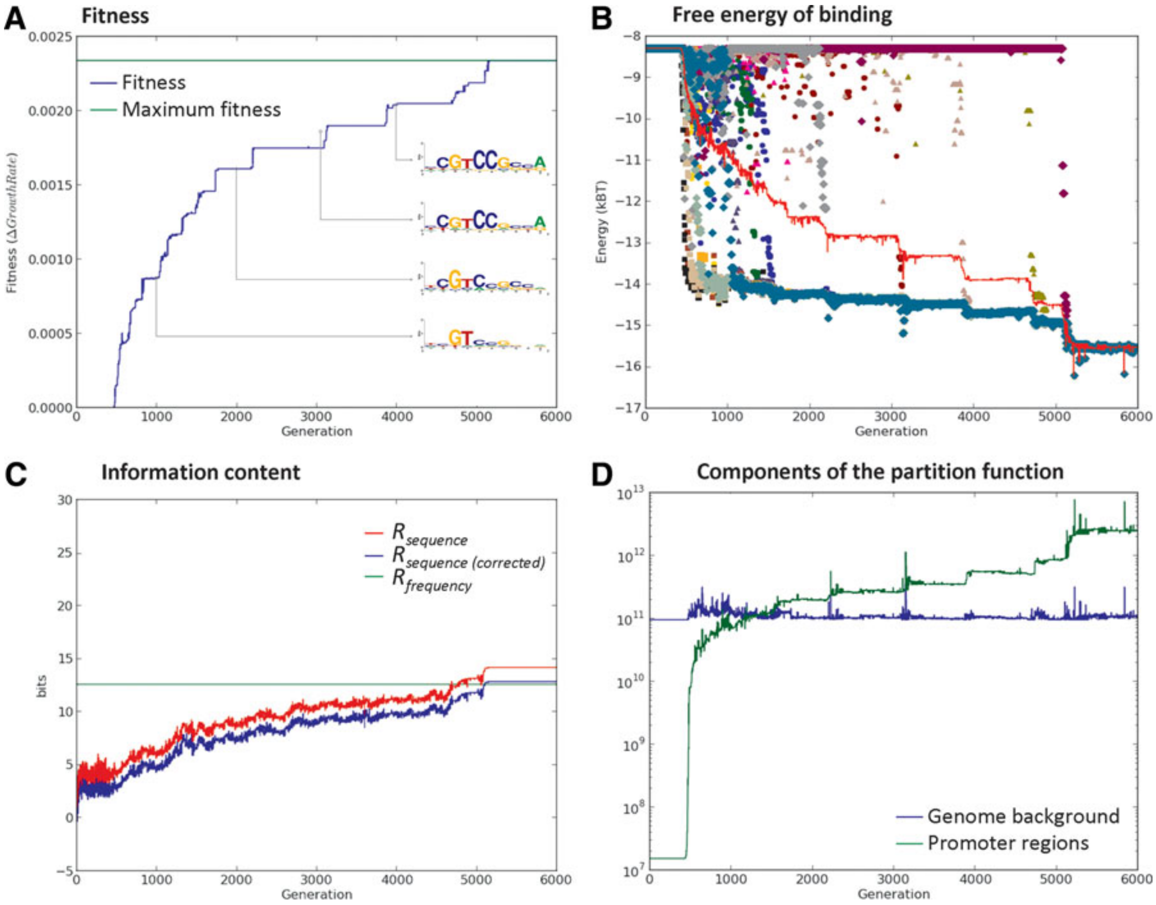

FIG. 2. Trace indicators for a complete evolutionary simulation. All values correspond to the fittest organism in the population at any given generation. **(A)** Organism fitness. Sequence logos illustrate the step-wise evolution of the associated TF-binding motif as new sites are recognized. **(B)** Free energy of binding for each site in the motif. **(C)** Information content R_{sequence} (corrected and uncorrected for small-sample error) of the evolved TF-binding motif. The predicted $R_{\text{frequency}}$ for the system is superimposed. **(D)** Contributions to the partition function from the sampled background (blue) and the promoter regions (green).

target sites, while maintaining the background at energies close to nonspecific binding. Thereafter, the simulation displays a step-wise increase in fitness that results from a progressive refinement of the recognizer and the gradual incorporation of additional sites. This behavior can also be traced by changes in the individual binding energies of sites (Fig. 2B) and by changes to the information content (R_{sequence}) of the associated TF-binding motif (Fig. 2C), and ultimately results in the TF assigning the preset fraction of probability mass to its target sites and the remainder to the background (Fig. 2D).

2.2. Evolutionary strategies for modulating regulatory activity

An important caveat of $R_{\text{frequency}}$ as a predictor of evolved information content for the purposes of evolutionary simulations is that it assumes a scenario that cannot be evolved in a realistic framework for a single-molecule model. The posterior entropy in $R_{\text{frequency}}$ assumes that all M sites targeted by the TF are bound with equal probability and that no other sites in the genome are bound, thereby assigning all the probability mass to M sites. Whereas these conditions can be approximated in nature by systems with multiple molecules of the TF, in a realistic simulation for the single-molecule case this ideal can never be achieved. In ESTReMo, where nonspecific binding is taken into account and where pseudo-sites cannot be completely eradicated by selection, the assumptions in $R_{\text{frequency}}$ result in the uninformative evolution of systems targeting motifs as close to the consensus as possible. To circumvent this problem, we sought to evolve systems in which only a fraction (α) of the total probability mass is assigned to target sites. This led us to explore possible generalizations of $R_{\text{frequency}}$ to this general case. Such an extension is bound by two opposing scenarios. If we assume that a fraction α of the total probability mass is assigned uniformly to M sites, the remaining $(1-\alpha)$ fraction can be distributed evenly over the genomic background (by extension of the maximum entropy assumption in $R_{\text{frequency}}$):

$$\begin{aligned} R^{*,\min}(\alpha) &= \left[-\sum_{i=1}^{G}\frac{1}{G}\log_2\frac{1}{G}\right]-\left[-\sum_{i=1}^{M}\frac{\alpha}{M}\log_2\frac{\alpha}{M}-\sum_{i=1}^{G-M}\frac{1-\alpha}{G-M}\log_2\frac{1-\alpha}{G-M}\right] \\ &= \log_2 G+\alpha\log_2\frac{\alpha}{M}+(1-\alpha)\log_2\frac{1-\alpha}{G-M} \end{aligned} \tag{4}$$

or it can be concentrated at a single genomic position:

$$\begin{aligned} R^{*,\max}(\alpha) &= \left[-\sum_{i=1}^{G}\frac{1}{G}\log_2\frac{1}{G}\right]-\left[-\sum_{i=1}^{M}\frac{\alpha}{M}\log_2\frac{\alpha}{M}-(1-\alpha)\log_2(1-\alpha)\right], \\ &= \log_2 G+\alpha\log_2\frac{\alpha}{M}+(1-\alpha)\log_2(1-\alpha) \end{aligned} \tag{5}$$

leading to a lower bound ($R^*_{\max}$) and an upper bound ($R^*_{\min}$) on the information content of a system capable of assigning a fraction α of its total probability mass uniformly among M sites.

In between these two bounds, a theoretical optimum may be defined as follows. On a uniform genomic background of size G, a classifier operating on R bits of information will recognize $2^{-R}\cdot G$ positives in expectation. The fraction α of its probability mass assigned to the M target sites is therefore:

$$\alpha=\frac{M}{2^{-R^{\dagger}}G}, \tag{6}$$

which leads directly to

$$R^{\dagger}(\alpha)=\log_2\frac{G}{M}+\log_2\alpha. \tag{7}$$

For any given α, the formula for $R^{\dagger}$ is therefore equivalent to $R_{\text{frequency}}$ for a TF targeting M/α sites, suggesting that the optimal strategy for adjusting the occupancy of target sites in a TF consists in exploiting a proportional number of pseudo-sites with similar affinity to the target sites. To analyze the general applicability of this principle, we ran simulations to evolve TFs capable of recognizing sixteen 16-bp-long binding sites on an effective background of 10,000 bp sampled from either the *Escherichia coli* K-12 or a randomly generated genome of equal length. Specifically, we monitored the evolved R_{sequence} for evolutionary runs in ESTReMo under increasing α values. The results (Fig. 3) reveal that $R^{\dagger}$ is a robust predictor of the evolved R_{sequence} (R^2 = 0.59, $p < 0.001$), validating this particular generalization of $R_{\text{frequency}}$. We

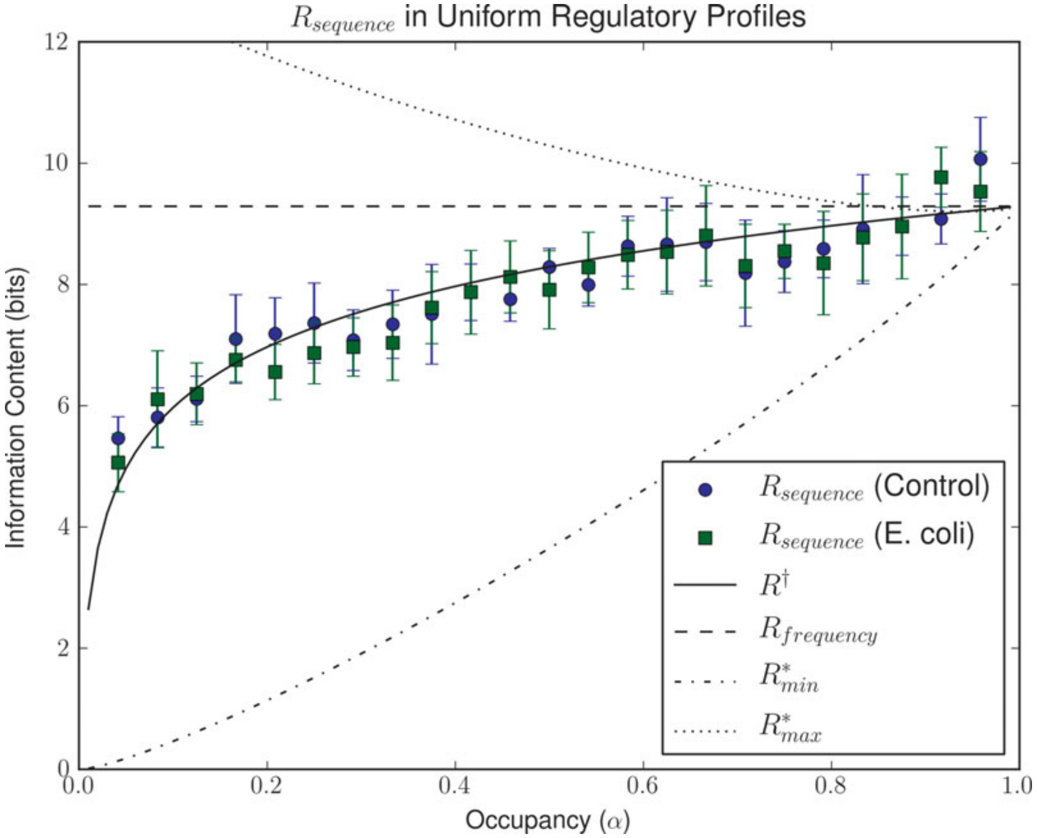

FIG. 3. Evolution of uniform regulatory systems. Plot shows evolved R_{sequence} values for simulations on randomly generated backgrounds (solid circles) and on the *Escherichia coli* genome (squares) as a function of the fraction of total probability mass (α) assigned to target sites. Error bars depict the 95% confidence interval for the mean. Corresponding values of $R_{\text{frequency}}$, R^{*}_{max}, R^{*}_{min}, and $R^{\dagger}$ are shown.

observe similar R_{sequence} values for simulations operating on *E. coli* and randomly generated uniform backgrounds (Mann–Whitney *U*-test $p > 0.05$). Despite this, TFs operating on real, larger genomic backgrounds may be able to exploit imbalances in the distribution of L-mers, already known to be targeted by restriction enzymes, to effectively decrease the informational demands of their associated TF-binding motifs (Gelfand and Koonin, 1997; Hahn et al., 2003).

The regulatory properties of a TF are a function of its binding affinity for its target sites, the genome it operates on, and the number of TF molecules present in the cell. Even though the formal derivation of $R^{\dagger}$ and the experimental results presented here apply only to the single-molecule case, they have obvious implications for the general case of multiple copy number. The binding of TFs to sites that do not directly regulate the expression of genes (i.e., pseudo-sites) has been traditionally interpreted as genetic noise faced by the TF and assumed to be indiscriminate following the random occurrence of pseudo-sites in large genomes (Berg and von Hippel, 1987; Robison et al., 1998). The evolutionary principle behind $R^{\dagger}$, however, suggests that TFs may be able to actively exploit pseudo-sites to modulate the effective occupancy of target regulatory sites, and that they may do so by relying on a small set of low-binding-energy pseudo-sites rather than on generic genome-wide sequestration by a full range of weak binding sequences. Hence, these results indicate that strong pseudo-sites may be considered functional and could be actively selected for by evolution.

2.3. Evolution of complex regulatory motifs

Transcriptional regulation of gene expression involves the precise tuning of both expression levels and temporal patterns of activation through the modulation of interactions between TFs, RNA–polymerase, and their respective binding sites (Browning and Busby, 2004; Friedman et al., 2005; Alon, 2007). Unfortunately, the precise effects of regulatory demands on the information content of evolved TF-binding motifs are hard to anticipate. Furthermore, the generalization of conventional approaches to predict information content in

TF-binding motifs, such as $R_{\text{frequency}}$, leads to counterintuitive results. The formula for $R_{\text{frequency}}$ in Equation 3 can be generalized to incorporate the entropy of a specific system that binds differentially to its target sites (as opposed to uniformly):

$$R_{diff} = -\sum_{G}\left(\frac{1}{G}\log_2\left(\frac{1}{G}\right)\right) - \left[-\sum_{i=1}^{M} p_i \log_2 (p_i)\right] \tag{8}$$

Because the posterior entropy is now smaller than in the standard $R_{\text{frequency}}$ case, the predicted information content must be higher. For instance, in a system targeting 16 sites on a 10,000 bp background, $R_{\text{frequency}}$ predicts that a recognizer needs 9.29 bits of information to correctly locate its sites. For a recognizer that binds 75% of its time to 4 of its 16 sites and only 25% of the time to the remaining 12, however, the generalized R_{diff} indicates that 10.8 bits are necessary. This same reasoning applies to $R^{\dagger}$ and other generalizations of $R_{\text{frequency}}$ and suggests that differential regulation imposes additional informational demands on TFs. It has been suggested that these informational requirements could be linked to guaranteeing accurate regulation throughout an extended regulatory range (Erill and O'Neill, 2009). Nonetheless, sites with decreased affinity for the TF should be more degenerate, from which it follows that a motif containing a significant number of ''weak'' sites should display lower information content (R_{sequence}).

We analyzed the effect of regulatory constraints on the evolution of TFs and their binding motifs by simulating the evolution of two illustrative types of regulatory systems targeting 16 sites on a 10,000 bp effective background. In bimodal systems, one half of the 16 sites required higher occupancy (α_{high}) than the other half (α_{low}). In stepwise regulatory ladders, each of the 16 sites was set to evolve a value of occupancy (α_i) that increased gradually from α_{low} to α_{high}. In both cases, we defined a reference value for total probability mass $\alpha = 0.5$, and we assigned sequential fractional values to α_{low} and α_{high} ($\alpha/4$, $\alpha/8$, and $\alpha/16$ for α_{low} and $\alpha \cdot 3/4$, $\alpha/2$, and $\alpha/4$ for α_{high}). The results (Figs. 4 and 5) reveal that the evolved

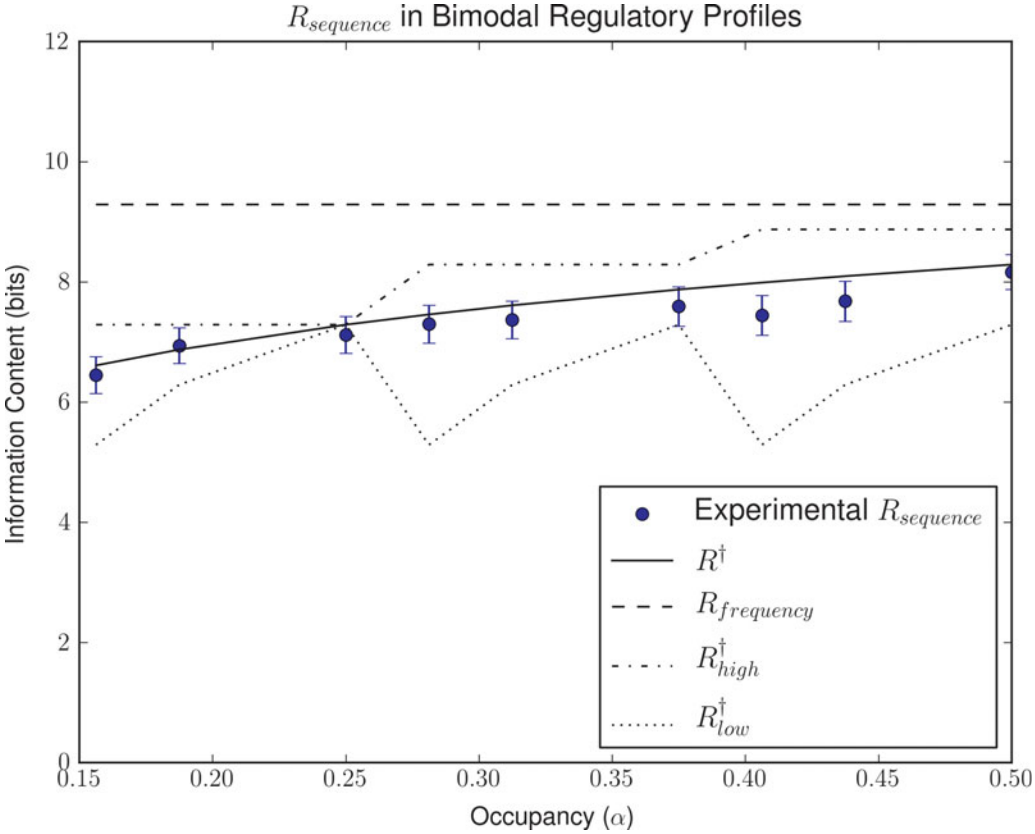

FIG. 4. Evolution of bimodal regulatory systems. The plot shows evolved R_{sequence} values for different evolutionary simulation scenarios with half of the sites set to higher occupancy (α_{high}) than the other half (α_{low}). Error bars depict the 95% confidence interval for the mean. For each R_{sequence} value, the corresponding $R^{\dagger}$ values for α_{high}, α_{low}, and α_{eff} values are shown.

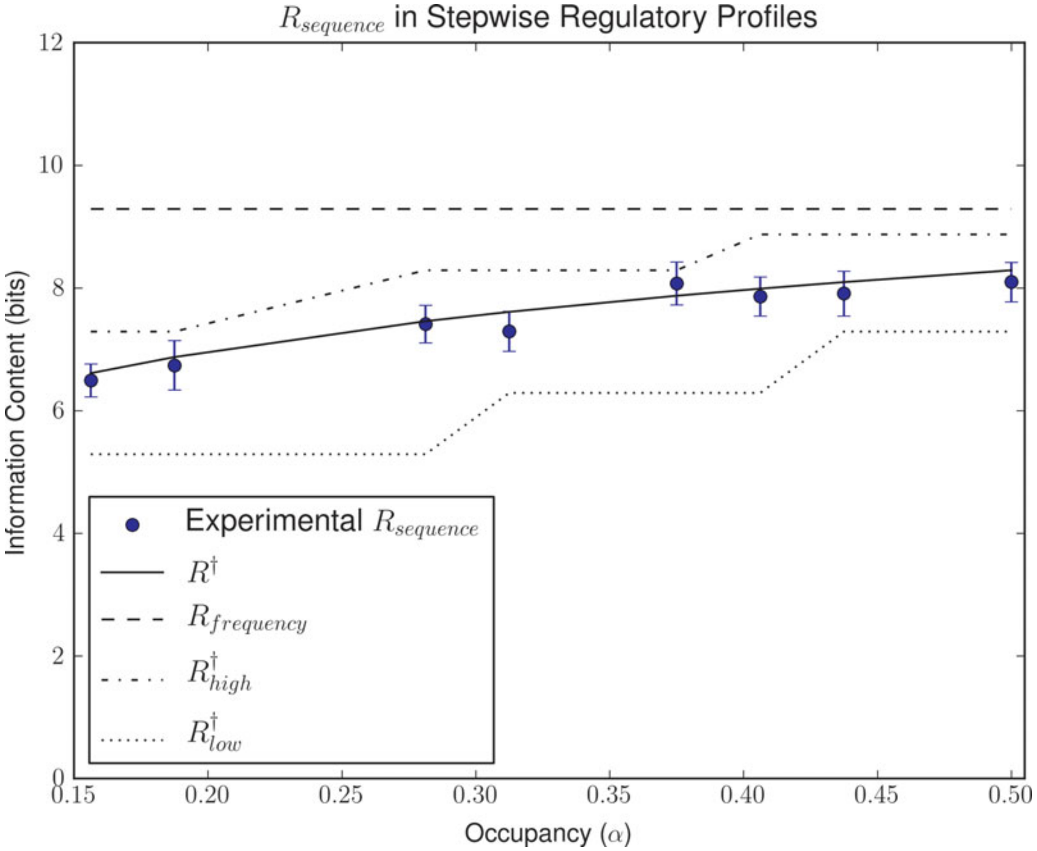

FIG. 5. Evolution of stepwise regulatory ladders. The plot shows evolved R_{sequence} values for different evolutionary simulation scenarios with sites gradually increasing from low (α_{low}) to high (α_{high}) occupancy. Error bars depict the 95% confidence interval for the mean. For each R_{sequence} value, the corresponding $R^\dagger$ values for α_{high}, α_{low}, and α_{eff} values are shown.

information content (R_{sequence}) traces $R^\dagger$ for the effective α (α_{eff}), defined as the total amount of probability mass assigned to all 16 sites in each setting ($R^2 = 0.23$, $p < 0.01$ for Fig. 4 and $R^2 = 0.22$, $p < 0.001$ for Fig. 5).

The results of bimodal and stepwise ladder systems strongly suggest that the information requirements of the system are dictated primarily by the effective fraction of total probability mass (α_{eff}) that the TF needs to assign to its target sites. In essence, the fraction of total probability mass assigned to target sites dictates the specificity of the TF, which imposes constraints on the required information content (R_{sequence}) of the TF-binding motif. As a result, the information content of a TF-binding motif for a system implementing an arbitrarily complex regulatory pattern can be estimated to first order by $R^\dagger(\alpha_{\text{eff}})$. To analyze the possible role of other contributing factors, such as the regulatory range that the TF needs to target, we conducted a series of simulations in which the regulatory targets for sixteen sites were sampled uniformly at random from the set of possible regulatory profiles. The results shown in Figure 6 reveal that the information content of evolved TF-binding motifs meeting these regulatory demands correlates with $R^\dagger(\alpha_{\text{eff}})$ ($R^2 = 0.66$, $p < 0.01$), and is effectively bound by $R^\dagger$ computed for 1 site at α_{high} and for 1 site at α_{low} (permutation test, $p < 0.01$). We detected no relationship, however, between information content and the coefficient of variation of the target occupancies ($R^2 = 0.004$, $p > 0.05$), suggesting that the dispersion of the regulatory targets is not explanatory after correction for mean occupancy. These results therefore support the previously stated hypothesis that the effective fraction of total probability mass is the major factor dictating the evolution of information content in TF-binding motifs. Furthermore, they show that for a given effective occupancy the span of required occupancies does not exact additional informational requirements on TF-binding motifs.

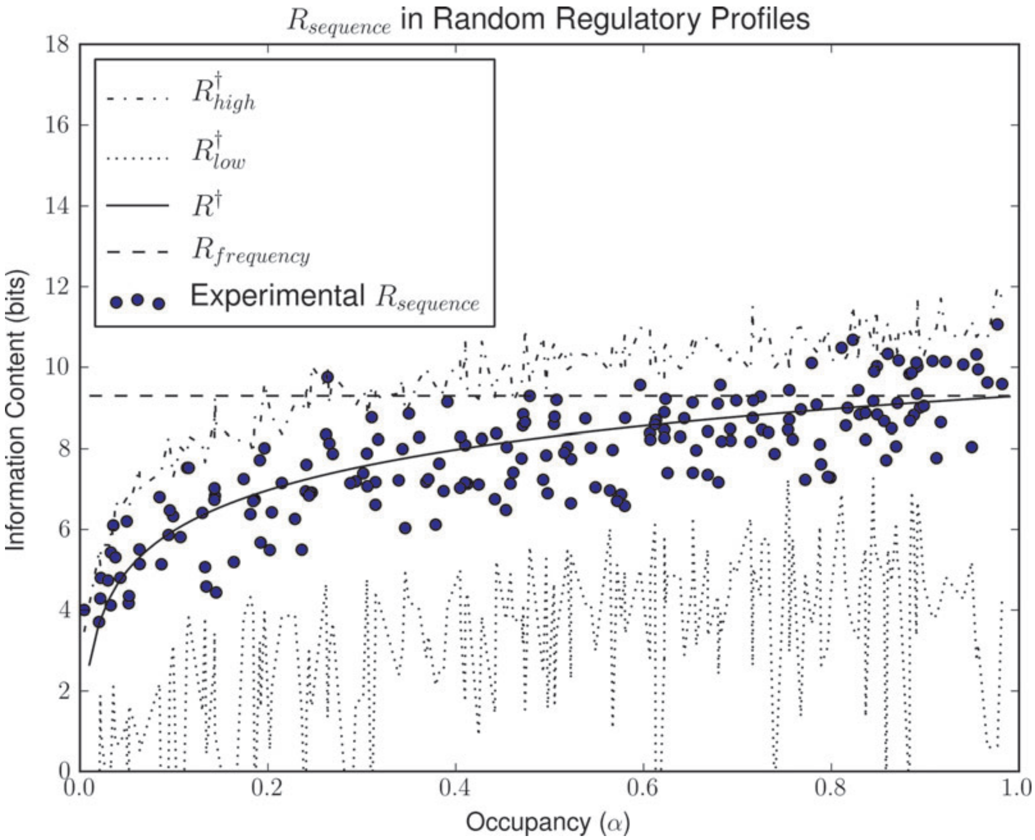

FIG. 6. Evolution of randomized regulatory scenarios. The plot shows 198 evolved R_{sequence} values for different evolutionary simulation scenarios. The full dataset is shown instead of summary statistics because of the uniqueness of α values. For each R_{sequence} value, the corresponding $R^{\dagger}$ values for α_{high} and α_{low} are shown.

3. CONCLUSIONS

The informational requirements for transcriptional systems with complex regulatory demands are not obvious. By generalizing $R_{\text{frequency}}$ into $R^{\dagger}$ as a predictive indicator for the evolved information content in TF-binding motifs and assessing its accuracy by means of evolutionary simulations, here we show that the informational requirements of arbitrarily complex regulatory patterns are dictated primarily by the effective fraction of probability mass assigned to target sites. We find that the presence of a large number of low-occupancy sites in a regulatory system can decrease the expected information content, but this effect is weak in comparison to the influence of the effective fraction of probability mass. As a result, the amount of information in TF-binding motifs for systems with complex regulatory demands can be effectively estimated to first order by $R^{\dagger}$.

4. METHODS

4.1. Evolutionary simulator

The evolutionary simulator used in this work, ESTReMo, has been developed integrally in C++ as an open-source project and is available for download online. ESTReMo implements a genetic algorithm with fixed population size, generational replacement, mutation, and *k*-tournament selection (Eiben and Smith, 2007). The genetic algorithm evolves a simulated TF (recognizer) and a set of promoter regions wherein TF-binding sites are expected to evolve. Promoter sequences are stored as DNA strings and encoded using a BIN4 notation for processing by the recognizer (Wu and McLarty, 2000). The recognizer is encoded as a set of weights that are instantiated into a predefined neural network architecture. We use a single-layer feed-forward network with linear activation function, which effectively maps to a position-specific weight-matrix model (Workman and Stormo, 2000). Independent mutation rates and operators are used for the recognizer and for the promoter regions. Mutations in promoter regions change the base at a given position following predefined nucleotide frequencies. Recognizer weights are mutated by adding or subtracting a factor (Δw) that is distributed as the cube of a standard uniform random variable. The output of the recognizer for the ith site S_i is interpreted as the change in free-energy (ε_i) associated with binding in units of K_BT, where K_B is the Boltzmann constant and T the absolute temperature. Occupancies for all positions in all promoter sequences are computed by assuming a Boltzmann distribution (Stormo and Fields, 1998):

$$P(S_i \text{ bound}) = \frac{e^{-\beta\varepsilon_i}}{Z}, \quad Z = \sum_{j\in\substack{\text{promoter}\\ \text{regions}}} e^{-\beta\varepsilon_j} + \sum_{j\in\substack{\text{genomic}\\ \text{background}}} e^{-\beta\varepsilon_j}, \tag{9}$$

where $\beta = 1/K_BT$ is inverse temperature and j is the genome coordinate. The genomic component Z_G of the partition function Z for the Boltzmann distribution is approximated by taking S contiguous samples (indexed by U_j) uniformly at random from a genome sequence and scaling their energy contributions to the desired effective genome size (G). The scaled partition function is then given by

$$\overline{Z}_G = \frac{G}{S}\sum_{j=1}^{s} e^{-\beta\varepsilon_{U_j}} \tag{10}$$

where ε_{Uj} is the energy associated with a given random sample U_j and S is the number of samples. For any given promoter sequence, expression of the associated gene is evaluated as the maximum occupancy for any site within the promoter sequence. Expression is assumed to encompass transcription and translation, resulting in a value of enzyme concentration within the cell. Target expression levels are defined as desired intracellular concentrations. The relationship between target expression level and fitness is assumed to follow the experimentally fitted cost–benefit function reported for LacZ (Dekel and Alon, 2005):

$$\text{fitness} = \delta\frac{EL}{K_Y + L} - \eta\frac{E}{1 - \frac{E}{M_E}} \tag{11}$$

where E is the enzyme (LacZ) concentration, L is the amount of lactose present in the medium (mM), K_Y is the Michaelis constant of LacY (mM), M_E is the maximum enzyme concentration, and δ and η are dimensionless

Table 1. Summary of Standard Parameters for the Simulations Reported in This Work

Category	*Parameter*	*Value*[a]
Genetic algorithm	Tournament size	2
	Promoter region mutation rate (μ_p)	0.001 mutations/bp/generation
	Recognizer mutation rate (μ_r)	0.005 mutations/weight/generation
	Population size	500
	Maximum number of generations	15,000
	Stopping criteria	Max generation & optimal fitness
	Recognizer weight initialization	0
	Promoter region base initialization	Uniform
Regulatory network	Promoter regions	16
	Size of promoter regions	32 bp
	Length of binding sites	16 bp
Genome	Size	10,000 bp
	Sequence	Random 50% GC/*E. coli*
Fitness model	Wild-type enzyme concentration	1 μM
	Benefit parameter (δ)	0.17
	Cost parameter (η)	0.02
	Maximum enzyme concentration (M)	1.8
	Michaelis constant of LacY (K_Y)	0.4 mM
Occupancy model	Distribution of transcription factor	Boltzmann
	Temperature	30°C
	Nonspecific binding	−8 kcal/mol
	Expression function	Max occupancy in promoter

[a] Variations to these parameters are explicitly noted in the text.

tuning parameters for benefit and cost, respectively. Equation 11 is defined in terms of wild-type concentrations for the enzyme (E, M_E). In practice, we assume an enzyme wild-type concentration of 1 μM (Lu et al., 2007; Flamholz et al., 2013). Values for all parameters in the fitness model are reported in Table 1.

Native ESTReMo C++ code can be compiled to run serially or in conjunction with a CUDA extension that offloads the estimation of the partition function to a General Purpose Graphical Processing Unit (GPGPU).

4.2. Evolutionary simulations

All simulations reported in this work targeted the evolution of a simulated TF capable of regulating sixteen 16-bp-long sites while facing a background genome with an effective size of 10,000 bp sampled from either the *E. coli* K-12 (NC_000913) or uniform random sequences of equivalent size. Only the forward strand was considered for the effects of the simulation. The mutation rate for promoter regions was set at 10^{-3} mutations per base per generation following previous work (Schneider, 2000), effectively setting the system slightly below the one mutation per genome per generation limit. The per-weight mutation rate for the recognizer then was adjusted empirically to a conservative rate below the limit at which simulations failed to converge. A complete list of parameters for the simulations is provided in Table 1. Simulations were run on a 2.80 GHz × 8 core Intel X58 system with 12 Gb of RAM and an NVidia 580 GPU unit, and on the University of Maryland Baltimore County (UMBC) High Performance Computing Facility Tara cluster. All simulations were run for 15,000 generations, and only simulations that attained a fitness value within 1% of the predicted maximum fitness were kept for analysis.

4.3. Data analysis

The ability of $R^{\dagger}$ to approximate the R_{sequence} values generated by the evolutionary simulation scenarios was assessed by computing the Pearson correlation coefficient of determination (R^2) and its associated *p*-value. Statistical support for the difference in R_{sequence} values evolved in real and randomly generated genomic backgrounds was assessed using a Mann–Whitney *U*-test. Support for bounds on evolved R_{sequence} values based on $R^{\dagger}$ computed at minimum (α_{low}) and maximum (α_{high}) occupancy values was assessed by a permutation test. All statistical tests were performed using the Python Scipy library. All code used to perform the data analysis may be found online at http://sourceforge.net/projects/estremo/.

ACKNOWLEDGMENTS

The authors would like to thank Joseph P. Cornish and Matthias K. Gobbert for contributions to the development of precursors to the ESTReMo platform. This work was supported by NSF Grant MCB-1158056. The UMBC High Performance Computing Facility is supported by the NSF MRI (CNS-0821258 and CNS-1228778) and SCREMS (DMS-0821311) programs and by the UMBC. We would also like to thank the four anonymous reviewers, whose insightful comments improved this article.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Alon, U. 2007. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461.

Badis, G., Berger, M.F., Philippakis, A.A., et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723.

Berg, J., Willmann, S., and Lässig, M. 2004. Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.* 4, 42.

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.

Browning, D.F., and Busby, S.J.W. 2004. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* 2, 57–65.

Dekel, E., and Alon, U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436, 588–592.

Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13, 2381–2390.

Eiben, A.E., and Smith, J.E. 2007. *Introduction to Evolutionary Computing*. Springer, Heidelberg.

Erill, I., and O'Neill, M.C. 2009. A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics* 10, 57.

Flamholz, A., Noora, E., Bar-Even, A., et al. 2013. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci. USA* 110, 10039–10044.

Friedman, N., Vardi, S., Ronen, M., et al. 2005. Precise temporal modulation in the response of the SOS DNA repair network in individual bacteria. *PLoS Biol.* 3, e238.

Gelfand, M.S., and Koonin, E.V. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25, 2430–2439.

Gerland, U., and Hwa, T. 2002. On the selection and evolution of regulatory DNA motifs. *J. Mol. Evol.* 55, 386–400.

Hahn, M.W., Stajich, J.E., and Wray, G.A. 2003. The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20, 901–906.

Huang, L., Guan, R.J., and Pardee, A.B. 1999. Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities. *Crit. Rev. Eukaryot. Gene Expr.* 9, 175–182.

Kim, J.T., Martinetz, T., and Polani, D. 2003. Bioinformatic principles underlying the information content of transcription factor binding sites. *J. Theor. Biol.* 220, 529–544.

Lu, P., Vogel, C., Wang, R., et al. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25, 117–124.

Maerkl, S.J., and Quake, S.R. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315, 233–237.

Mardis, E.R. 2007. ChIP-seq: welcome to the new frontier. *Nat. Methods* 4, 613–614.

Minchin, S.D., and Busby, S.J. 2009. Analysis of mechanisms of activation and repression at bacterial promoters. *Methods* 47, 6–12.

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

Orphanides, G., and Reinberg, D. 2002. A unified theory of gene expression. *Cell* 108, 439–451.

Ptashne, M. 2005. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem. Sci*. 30, 275–279.

Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* 284, 241–254.

Roider, H.G., Kanhere, A., Manke, T., et al. 2007. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23, 134–141.

Schneider, T.D. 2000. Evolution of biological information. *Nucleic Acids Res.* 28, 2794–2799.

Schneider, T.D., and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100.

Schneider, T.D., Stormo, G.D., Gold, L., et al. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–431.

Stewart, A.J., Hannenhalli, S., and Plotkin, J.B. 2012. Why transcription factor binding sites are ten nucleotides long. *Genetics* 192, 973–985.

Stormo, G.D., and Fields, D.S. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* 23, 109–113.

Workman, C.T., and Stormo, G.D. 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput*. 2000, 467–478.

Wu, C.H., and McLarty, J.M., eds. 2000. *Neural Networks and Genome Informatics*, 1st ed. Elsevier Science, Oxford.

Zhao, Y., and Stormo, G.D. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29, 480–483.

Address correspondence to:
Dr. Ivan Erill
Department of Biological Sciences
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250

E-mail: erill@umbc.edu