Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites

Ville Mustonen*, Justin Kinney[†], Curtis G. Callan, Jr.[†‡§], and Michael Lässig*[§]

*Institut für Theoretische Physik, Universität zu Köln, Zülpicherstrasse 77, 50937 Köln, Germany; and [†]Joseph Henry Laboratories and [‡]Princeton Center for Theoretical Physics, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., June 19, 2008 (sent for review February 27, 2008)

We present a genomewide cross-species analysis of regulation for broad-acting transcription factors in yeast. Our model for binding site evolution is founded on biophysics: the binding energy between transcription factor and site is a quantitative phenotype of regulatory function, and selection is given by a fitness landscape that depends on this phenotype. The model quantifies conservation, as well as loss and gain, of functional binding sites in a coherent way. Its predictions are supported by direct cross-species comparison between four yeast species. We find ubiquitous compensatory mutations within functional sites, such that the energy phenotype and the function of a site evolve in a significantly more constrained way than does its sequence. We also find evidence for substantial evolution of regulatory function involving point mutations as well as sequence insertions and deletions within binding sites. Genes lose their regulatory link to a given transcription factor at a rate similar to the neutral point mutation rate, from which we infer a moderate average fitness advantage of functional over nonfunctional sites. In a wider context, this study provides an example of inference of selection acting on a quantitative molecular trait.

binding energy | transcriptional regulation | quantitative molecular trait

Regulatory elements can often be distinguished from background sequence by their evolutionary conservation. At the same time, it has become clear that many regulatory functions are not widely conserved, but are specific to certain species or clades (1). Thus, it seems likely that the evolution of regulatory function and, in particular, of *cis*-regulatory elements is a key component in evolutionary innovation and the differentiation between species (2–7). However, functional changes in regulation are difficult to gauge from sequence divergence alone. For example, many different sequence states of a promoter may lead to similar binding of transcription factors and thus have similar effects on the transcription of a regulated gene. Thus, discerning function from sequence requires a *phenotype* for regulatory elements and an evolutionary model to quantify natural selection acting on this phenotype.

In this article we address regulatory evolution from a biophysical perspective. We show that the binding energy of a transcription factor (TF) provides a quantitative phenotype for its target sites, and we develop a predictive model for binding site evolution based on this phenotype. A key ingredient of this model is the mapping from genotype to phenotype, that is, the sequence dependence of the binding energy for a given TF. Direct energy measurements are available for a few (mostly prokaryotic) transcription factors (8, 9), but low-throughput experiments generally do not provide enough data to fully constrain the energy function. By contrast, high-throughput binding assays (10, 11) provide copious, if indirect, data on TF binding to promoter regions, and we use recently developed methods to infer binding energies from such data for broad-acting TFs, such as yeast Abf1 (12, 13). Moreover, a high-throughput method of measuring TF binding energy to many sequences at once (14) promises to produce reliable energy data for a wider range of TFs. These energies define a phenotype in a population-genetic model of binding site evolution that contains genetic drift, point mutations, and selection acting on this phenotype (15–17). The average selective advantage of a site with given energy over mutants with different energy defines a *fitness landscape* (18), which can be inferred from energy histograms of functional sites and background sequence as shown in ref. 16. In the spirit of this model, we identify functional sites by evolutionary conservation of energy rather than sequence. This link between biophysics and evolution is the main difference between our approach and previous population-genetic studies (19, 20) of yeast TF binding site evolution, where selection acting on functional sites is defined and inferred at the level of individual nucleotides.

The fitness landscape we derive for yeast Abf1 binding sites displays a strikingly nonlinear dependence on energy. Stochastic simulations of binding site evolution in this fitness landscape are in quantitative agreement with a genomewide cross-species comparison between four yeast species: *Saccharomyces cerevisiae* (*cer*), *S. paradoxus* (*par*), *S. mikatae* (*mik*), and *S. bayanus* (*bay*) (see Fig. 1 for the phylogenetic tree and ref. 21 for the sequences of these species).

In particular, we find ubiquitous compensatory mutations within functional sites, such that the energy phenotype evolves under stronger constraint than does sequence. The compensations are due to the nonlinearity of the fitness landscape, which leads to fitness interactions (epistasis) between any two nucleotides of the target sequence, even if such interactions are absent in the sequence-dependent energy itself. By cross-species comparison, we can also quantify gain or loss of function between promoters of orthologous genes. We find that such events occur at a rate smaller but similar to the neutral point mutation rate, in general agreement with results of ref. 20. Evolution of promoter function can be integrated in our model if typical functional sites involve a moderate fitness advantage over nonfunctional sites; in that case, loss and gain of function can take place by mutations and genetic drift, without requiring changes in selection itself. The consequences of these findings for our conceptual picture of regulatory networks and for the bioinformatic analysis of regulatory sequences are considered under *Discussion*.

Results

Sequence-Dependent Binding Energy Models. We assume that the binding energy of a TF to a site with sequence $(a_1, \dots, a_\ell)$ is

Author contributions: V.M., J.K., C.G.C., and M.L. designed research; V.M., J.K., C.G.C., and M.L. performed research; V.M., J.K., and C.G.C. analyzed data; and V.M., C.G.C., and M.L. wrote the paper.

The authors declare no conflict of interest.

[§]To whom correspondence may be addressed. E-mail: ccallan@princeton.edu or lassig@thp.uni-koeln.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0805909105/DCSupplemental.

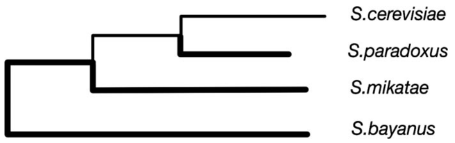

Fig. 1. Phylogeny of the four yeast species *S. cerevisiae* (*cer*), *S. paradoxus* (*par*), *S. mikatae* (*mik*), and *S. bayanus* (*bay*). We study the statistics of functional binding sites conserved in all four species, as well as functional changes on any of the external branches to *par*, *mik*, and *bay* (shown as thick lines). Figure adapted from ref. 21.

given by a 4 × ℓ matrix describing the additive energy contribution of each possible nucleotide a at each site position i,

$$E = \sum_{i=1}^{\ell} \varepsilon_i(a_i). \qquad [1]$$

In the absence of interactions with other factors, the binding probability of a TF to a site is governed by the equilibrium thermodynamics expression $p(E) = 1/[1 + \exp(E - \mu)/k_B T]$, where μ is the chemical potential of the TF in the cytoplasm. As a result, sites with energy $E < \mu$ are likely to be occupied, and sites with energy $E > \mu$ are unlikely to be occupied (22). Hence, we can infer model parameters μ and $\varepsilon_i(a)$, which give the best match between the pattern of promoter binding observed in PBM and ChIP-chip experiments (10, 11) and predicted site occupation (12, 13). We will focus on Abf1 and Reb1 in yeast as examples of TFs with many binding sites (for Reb1 results, see supporting information (SI) *Text*, Figs. S1–S2). Such TFs are well suited to our evolutionary analysis: a large ensemble of sites allows accurate inference of energy matrices and of site evolution statistics. Because the evolution of the relevant parts of the Abf1 protein is highly constrained [Abf1 in *cer* and *bay* differ only in 19 of 226 amino acids implicated in DNA binding (23)], we will assume that the evolution of regulatory interactions of Abf1 effectively reduces to that of its target sites. We describe Abf1 binding by an energy matrix of length $\ell = 14$ (which includes all highly specific core positions) inferred by the methods of ref. 12 from the genomewide ChIP-chip assay data of ref. 10. In the following, E denotes the (positive) energy difference between the actual sequence and the strongest binding sequence, measured in units of the chemical potential μ (supplementary measurements would be needed to determine the absolute value of μ).

Inference of Functional Sites by Energy Conservation. Scanning the intergenic sequence of *cer* with the Abf1 energy matrix produces the histogram of energy counts shown in Fig. 2*A*. For $E > 1.2$ the distribution is close to Gaussian and is well fit by the normalized energy distribution $P_0(E)$ resulting from uncorrelated random site sequence, with single-nucleotide frequencies $p_0(\mathrm{a})$ chosen to match intergenic averages. The low-energy excess of >500 counts above the P_0 distribution provides a rough estimate of the number of functional Abf1 sites, defined as loci where Abf1 binding is favored by natural selection. The energy histogram of excess sites then gives an estimate of the normalized energy distribution $Q_f(E)$ of functional sites. However, an energy cutoff $E < \widetilde{E}$ in a single species is not adequate to discriminate individual functional from nonfunctional sites: The distributions $Q_f(E)$ and $P_0(E)$ overlap, and any set Q_c of *cer* sites with $E < \widetilde{E}$ has a fraction of background sites that increases rapidly with $\widetilde{E}$ (16, 24). Evolutionary conservation can be used to prune out these false positives (19, 25–31). Here, we use conservation of the energy phenotype to generate the ensembles of functional sites for our subsequent analysis. Let Q_{cpm} be the set of triplets of orthologous sites (see *Methods*) in *cer*, *par* and *mik* which have $E < \widetilde{E}$ in all three species, and define similar sets Q_{cpb} and Q_{cmb} (using an obvious extension of notation). Choosing $\widetilde{E} = 0.9$, these sets contain between 727 and 791 sites, while the subset Q_4, with $E < \widetilde{E}$ in all four species, contains 708 members. The energy distributions of these sets provide estimates of $Q_f(E)$, which are quite similar to each other and to the single-species low-energy excess counts in *cer*. The remaining fraction of false positives caused by neutrally evolving background loci and loci with sequence constraint unrelated to Abf1 binding is estimated to be, at most, a few percent (see *SI Text*, Fig. S3). We conclude (*i*) that most of the sites contained in the Q sets are functional Abf1 binding sites and (*ii*) that most functional sites are conserved between all four species, with a stationary phenotype distribution $Q_f(E)$. Differences between the three-species sets will be used below to infer loss-of-function events.

In a second step, we identify Abf1 sites overlapping with other functional sequences. Such overlaps are known to be common in yeast (29, 32) and would confound our inference of selection for Abf1 binding. Specifically, we partition the set Q_4 into a subset Q_4^{ol} of 347 sites overlapping (by any amount) with another site predicted by using the comprehensive list of yeast TF motifs in ref. 27 (see *SI Text*), and a complementary subset Q_4^{no} of 361 nonoverlapping sites. The energy distributions of both sets of sites are very similar; we use that of Q_4^{no} to infer $Q_f(E)$ for our evolutionary analysis. Direct comparison with binding assay data provides a useful consistency check: we find that promoters containing Q_4^{no} sites are strongly biased toward high experimental binding assay scores in both ChIP-chip (10) and PBM (11) data; see Fig. 2*B* and Fig. S4.

Fitness Landscape for Functional Sites. The different phenotype distributions $Q_f(E)$ and $P_0(E)$ result from differences in the

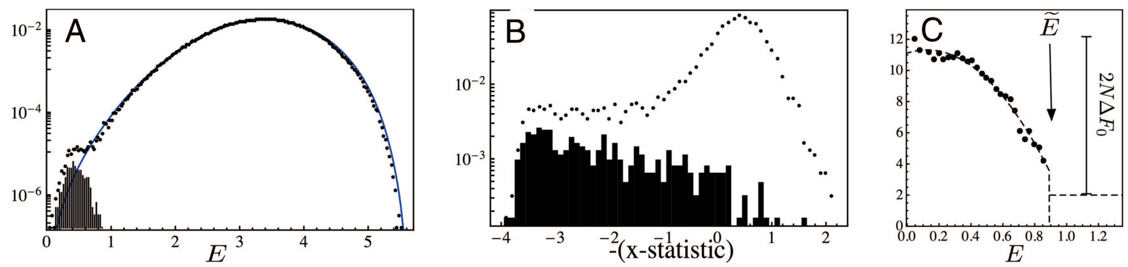

Fig. 2. Abf1 binding energy and binding assay distributions. (*A*) Histograms of site energies, as predicted by the Abf1 energy matrix for all *cer* intergenic sequence (dots), for neutral background sequence described by the distribution $P_0(E)$ (line), and for the 361 Abf1 binding sites in the Q_4^{no} ensemble described by the distribution $Q_f(E)$ (bars). (*B*) Comparison with ChIP-chip data of ref. 10: histogram of log intensities x of *cer* binding for all promoters (dots) and for the subset of promoters containing Q_4^{no} sites (bars), which are mostly in the tail of the distribution (≈80% in the regime < -1). (*C*) Fitness landscape for Abf1 binding sites. Binding range ($E < \tilde{E} = 0.9$): scaled landscape $2N\,F(E) = \log[Q_f(E)/P_0(E)]$ by using energy data from Q_4^{no} functional sites and intergenic background sequence (dots), quadratic fit (dashed line). Nonbinding range ($E > \tilde{E}$): $2N\,F(E)$ is approximated as constant with difference $2N\,\Delta F_0$ to maximal binding.

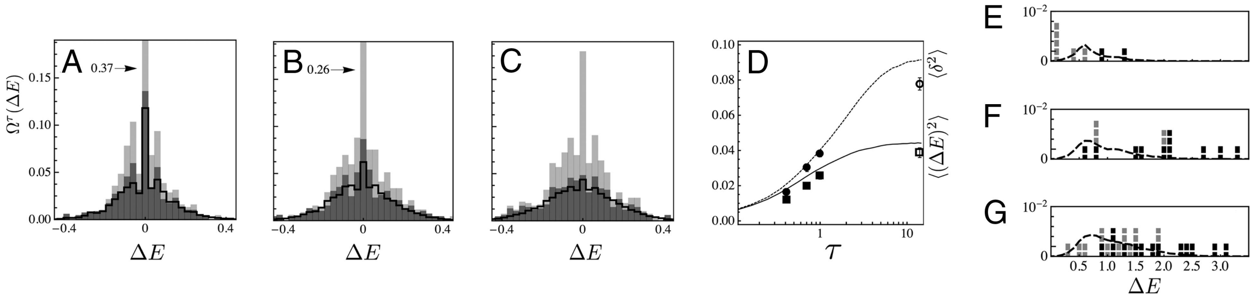

Fig. 3. Phenotype evolution of binding sites. (*A*) Histogram of *cer*–*par* energy differences $\Delta E = E_{par} - E_{cer}$ for conserved Abf1 binding sites in $\mathcal{Q}_4$ (bars), giving the sum of counts from nonoverlapping sites (dark shaded part) and overlapping sites (light shaded part). Predicted distribution Ω^τ (ΔE), normalized to total number of nonoverlapping sites (solid line). (*B* and *C*) Same as *A* for *cer*--*mik* and *cer*–*bay* energy differences. (*D*) Energy divergence $\langle(\Delta E)^2\rangle$ (filled squares) and additive divergence δ^2 (dots, see text) for conserved nonoverlapping sites between *cer* and the other three species, plotted against evolutionary distance τ; predicted energy divergence $\langle(\Delta E)^2\rangle$ (τ) (solid line) and additive divergence $\delta^2(\tau)$ (dashed line). The large-τ limit of these functions reproduces (up to sampling effects) the site-to-site energy variance in *cer* (open square) by definition of the evolution model, and predicts the site-to-site linear variance (open dot). (*E*) Histogram of energy differences $\Delta E = E_{par} - E_{cer}$ for the seven three-species conserved sites in $\mathcal{Q}_{cmb}$ without functional ortholog in *par* (bars), events with indels are highlighted (black bars). Theoretical distribution of energy changes $\omega\Omega^\tau_{\text{loss}}$ (ΔE) for loss events caused by point mutations (dashed line). (*F* and *G*) Same for species-specific loss of function in *mik* (13 events) and in *bay* (23 events).

evolutionary dynamics between functional loci and genomic background. At a background locus, neutral point mutations, insertions, and deletions change one sequence state $\mathbf{a} = (a_1, \ldots, a_\ell)$ into another state $\mathbf{b}$ with substitution rates $\mu_{\mathbf{a}\to\mathbf{b}}$. At a functional locus, each sequence state $\mathbf{a}$ has a Malthusian fitness, which we assume depends only on its energy phenotype: $F(\mathbf{a}) = F(E(\mathbf{a}))$. Substitutions at a functional locus take place at a different rate $u_{\mathbf{a}\to\mathbf{b}}$, which depend on the selection coefficient $\Delta F = F(\mathbf{b}) - F(\mathbf{a})$ and on the effective population size N as given by the classic Kimura–Ohta theory (33–35) (see *Methods*). Therefore, if neutral evolution leads to an equilibrium distribution $P_0(\mathbf{a})$ of sequence states, evolution under selection will generate a different distribution $Q(\mathbf{a})$. The same is true for the resulting phenotype distributions $P_0(E)$ and $Q(E)$, the log ratio of which reproduces the fitness landscape scaled by a factor $2N$ (15, 16):

$$2NF(E) = \log\left[\frac{Q(E)}{P_0(E)}\right] + \text{const.}, \qquad [2]$$

where the constant is arbitrary, because only fitness differences matter. The equilibrium distributions $Q(E)$ and $P_0(E)$ are defined by the sequences of an *individual* locus over time, which are not accessible to observation. To infer the fitness landscape from Eq. **2**, we replace these distributions by *ensemble* distributions generated by all functional or background loci for a given transcription factor in a given genome. Specifically, the distributions $Q_f(E)$ and $P_0(E)$ derived in the previous paragraph will be used to define the fitness function $F(E)$ of Abf1 in *cer* for $E < \widetilde{E}$ (dashed line in Fig. 2*C*).

Our inference of an energy-dependent fitness landscape relies on the uniformity of selection in time and across loci. First, we assume that the substitution dynamics of an *individual locus* has relaxed to equilibrium, which requires that its fitness function has remained approximately constant for at least the relevant equilibration time. This is plausible, because we do not require equilibrium at the level of site sequence, which would be approached very slowly, but only at the level of energy phenotype. Under this weaker condition, the equilibration time is the time for adaptive formation of a binding site, which can be much shorter than the neutral mutational time scale $1/\mu_0$ (15). Deviations from equilibrium would not invalidate our inference of selection: the equilibrium method produces a lower bound to the actual level of selection, as shown by model studies of time-dependent selection (36). We note that the phenotypic equilibrium assumption is consistent with the observation that the phenotype distributions $Q(E)$ and $P_0(E)$ [and the inferred fitness landscapes $F(E)$] are remarkably similar in the four species of this study, despite their very different divergence times (Fig. S5). Inferring these distributions involves averaging *across loci* in one genome, a procedure analogous to the standard derivation of a position weight-scoring matrix from a set of functional loci. This inference contains a second assumption: selection must be sufficiently uniform across loci so that the evolutionary information in the average fitness function is not overwhelmed by differences between sites, which arise from their sequence context (for example, from cooperative binding with neighboring sites), the function of the regulated gene, and a number of other factors. We have at present no way to infer site-specific fitness functions. However, at least for the broad-acting factors of this study, the average fitness landscape $F(E)$ makes remarkably accurate predictions of binding site evolution, providing *a posteriori* evidence for the uniformity assumption.

Phenotype Evolution of Conserved Sites. We can predict the evolution of sequence and phenotype for functional binding sites by stochastic simulation of the Kimura–Ohta point substitution process with rates $u_{\mathbf{a}\to\mathbf{b}}$ specified the corresponding neutral rates $\mu_{\mathbf{a}\to\mathbf{b}}$ and by the fitness landscape $F(E)$ (see *Methods* and *SI Text*). Insertions and deletions play a minor role for conserved sites, but are important for loss of function as discussed below.

For fixed evolutionary distance τ, the stochastic simulation generates an ensemble of orthologous sequence pairs; the distribution of their energy differences ΔE, which we denote by $\Omega^\tau_f(\Delta E)$, is a useful way of quantifying phenotype evolution. Fig. 3 *A*–*C* shows the predicted ΔE distributions for three different values of τ, corresponding to the evolutionary distance between *cer* and each of the other three species. In these figures, histograms of cross-species phenotypic differences for the set $\mathcal{Q}_4$ of conserved functional Abf1 sites are shown separately for the subsets $\mathcal{Q}_4^{\text{no}}$ and $\mathcal{Q}_4^{\text{ol}}$. For the putatively single-function sites $\mathcal{Q}_4^{\text{no}}$, the predicted distributions (solid lines) agree closely with the observed distributions (dark-gray histograms), indicating that the fitness landscape $F(E)$ adequately describes selection for the function of Abf1 binding. A more compact measure of phenotypic divergence is given by the mean square site energy difference, $\langle(\Delta E)^2\rangle$, between *cer* and any of the other three species. The increase of phenotypic divergence with evolutionary distance for the $\mathcal{Q}_4^{\text{no}}$ set is displayed in Fig. 3*D*. The observed values are in good agreement with the theoretical predictions derived from the distribution $\Omega^\tau_f(\Delta E)$. In the limit of large τ, evolutionary correlations are lost, and the variance of $\Omega^\tau_f(\Delta E)$ reproduces the variance of the single-species distribution $Q_{\mathrm{f}}(E)$ (see also Fig. 3*D*).

The set Q_4^{ol} of overlapping sites, which contain putative multifunctional sequences, evolve under much stronger constraint than nonoverlapping sites, as shown by the light-gray part of histograms in Fig. 3 *A–C*. Many of these sites are almost completely conserved in their sequence, which implies complete phenotypic conservation ($\Delta E = 0$, see the corresponding peaks in the histograms). Remarkably, the energy distribution of the subset Q_4^{ol} is still given by $Q_{\mathrm{f}}(E)$, as shown in Fig. S6*a*. This indicates that overlapping sites are still functional Abf1 sites and selection for Abf1 binding is still described by the same fitness landscape $F(E)$, whereas selection for the additional regulatory functions acts independently on other phenotypes. The joint selection for more than one function introduces substantial negative selection on most nucleotide substitutions, and thus drastically cuts the sequence space accessible by near-neutral and compensatory changes in the fitness landscape $F(E)$ for a single factor.

Epistasis and Compensatory Mutations. A striking feature of the Abf1 fitness landscape shown in Fig. 2*C* is its nonlinear form. The negative curvature of $F(E)$ in the binding regime ($E < \widetilde{E}$) expresses the evolutionary coupling between different nucleotides within a binding site by fitness interactions (epistasis): a nucleotide change at a given position of the site induces a change of binding energy $\Delta E = E_2 - E_1$, which is independent of nucleotides at other positions in our energy model; however, the selection coefficient $\Delta F = F(E_2) - \mathrm{F}(E_1)$ of this change depends on the initial energy E_1 and therefore on all other nucleotides within the site. In a linear fitness landscape, by contrast, ΔF would be proportional to the phenotype change ΔE and would be independent of the nonmutating nucleotides.

Epistasis effects are clearly visible in the evolution of conserved binding sites. The energy difference between orthologous sites is a sum of contributions from base substitutions at individual positions: $\Delta E = \Sigma_{i=1}^{\ell} \Delta\varepsilon_{\mathrm{i}}$. We now compare the energy divergence $\langle(\Delta E)^2\rangle$ for nonoverlapping conserved sites with the additive divergence $\delta^2 \equiv \Sigma_{i=1}^{\ell} \langle(\Delta\varepsilon_i)^2\rangle$ evaluated for the same set of sites. Clearly, if substitutions within one site were independent of each other, the two divergence measures would be equal. However, Fig. 3*D* shows that $\langle(\Delta E)^2\rangle$ is significantly smaller than δ^2 for the actual evolution process of Q_4^{no} sites. This additional constraint on binding energy can be shown to be *independent* of the cutoff $\widetilde{E}$ used to define conserved sites (see *SI Text*, Fig. S3*b–e*). This is evidence that the effect is indeed caused by epistasis of the fitness landscape and is not an artifact of the cutoff introduced in our definition of conserved sites. The implication is that mutations are often compensatory: a substitution that decreases factor binding, which is deleterious on average, will increase the likelihood of a subsequent advantageous change restoring binding. Epistasis thus generates additional constraint on phenotype and function. This effect increases with evolutionary distance and is most pronounced in the large-distance limit, that is, for mean square differences between sites within one species. It cannot be captured by any linear fitness landscape. Previous evolutionary studies of yeast binding sites (19, 20) assume that the fitness of a site is linear in the contribution of its nucleotides, $F_{\mathrm{lin}}\ (\mathbf{a}) = \Sigma_{i=1}^{\ell} f_{\mathrm{i}}(a_i)$, which amounts to assuming that substitutions at different positions are independent of each other. The fitness functions $f_{\mathrm{i}}(\mathrm{a})$ are inferred from the position-dependent nucleotide frequencies $q_{\mathrm{i}}(\mathrm{a})$ at functional sites and the background nucleotide frequencies $p_0(\mathrm{a})$ by $2N f_i(a) = \log[q_i(a)/p_0(a)]$, an expression similar to Eq. **2** (37). This method correctly captures selective constraint at the single-nucleotide level but neglects compensation between mutations. It necessarily predicts that $\langle(\Delta E)^2\rangle_{\mathrm{lin}} = \delta^2$, underestimating the actual constraint on phenotype and function.

Loss and Gain of Function. Imposing the phenotype conservation criterion $E < \widetilde{E}$ in the three species *cer*, *par*, and *mik*, but not in *bay*, produces a list of 23 sites which have a nonfunctional ($E > \widetilde{E}$) orthologous site in *bay*, with no other functional Abf1 site present in the same *cis*-regulatory region of *bay* (see *Methods*). Most of these 23 sites appear to be functional Abf1 sites in *cer*, *par*, and *mik*: their energy histogram is consistent with the functional site distribution $Q_{\mathrm{f}}(E)$, and not with the background distribution $P_0(E)$; they also show a similar correlation with the ChIP-chip data as do the Q_4^{no} sites. Thus, the most likely explanation is that the regulation of these genes has changed after speciation from the last common ancestor of all four species: it has either been lost on the phylogenetic branch to the outgroup species *bay* or gained on the branch to the last common ancestor of the ingroup *cer*, *par*, and *mik* (which is less likely because the branch is shorter; see Fig. 1). Proceeding in a similar way for the sets Q_{cpb}^{no} and Q_{cmb}^{no}, we find 13 species-specific loss-of-function events in *mik*, and 7 in *par*. These numbers translate into an average loss rate of Abf1 regulatory sites of $\mu_{\mathrm{f}} \sim 0.1\ \mu_0$ (where μ_0 is the neutral point mutation rate).

Understanding functional *cis*-regulatory changes in the framework of our population-genetic model requires extending the fitness landscape $F(E)$ into the nonbinding range $E > \widetilde{E}$. Because energy changes in this range have only a small effect on the expression of the regulated gene, they are near-neutral and we can approximate the fitness landscape for $E > \widetilde{E}$ by a constant, as shown in Fig. 2*C*. This extension involves a single additional parameter, namely the fitness difference ΔF_0 between a nonfunctional site and a strong binding site. The value of ΔF_0 can be estimated by simulating the evolution of initially functional sites in the extended fitness profile and fitting the loss rate to our cross-species observations. We obtain a moderate level of selection $2N\Delta F_0 = [6 - 9]$ (the lower bound uses all loss events as defined in *Methods*, and the upper bound, the ones with only point mutations). This is comparable to the selection pressure previously found for Crp binding sites in *Escherichia coli* (16). Thus, conservation of sites is not complete: functional binding sites in the last common ancestor remain functional in all four species in 65–94% of the cases. A recent experimental study finds even faster divergence of regulatory function: only ≈20% of Ste12 and Tec1 binding events are conserved between *cer*, *mik*, and *bay* (7).

Phenotype changes between orthologous sites are significantly larger for loss and gain events than for conserved functional sites, as shown by the histogram of energy differences between functional sites in *cer* and their nonfunctional orthologs in one of the other three species (Fig. 3 *E–G*). The data are qualitatively reproduced by the predicted distribution $\Omega_{\mathrm{loss}}^{\tau}\ (\Delta E)$ of energy differences for loss events obtained by simulation of the point substitution process in the full fitness landscape $F(E)$. The full distribution of energy differences between functional sites and their orthologs is the weighted sum $(1 - \omega)\Omega_f^{\tau}(\Delta E) + \omega\Omega_{\mathrm{loss}}^{\tau}(\Delta E)$, where ω is the fraction of lost sites. Its bimodality is again due to the nonlinearity of the fitness landscape. Any linear function $F(E)$ would generate a unimodal $\Omega(\Delta E)$ and would not make a distinction between the phenotype statistics of conservation and functional change. As shown in Fig. 3 *E–G*, the average energy difference $\langle \Delta E \rangle$ of observed loss and gain events is somewhat larger than predicted. This is probably mainly because of sequence insertion and deletion processes (indels), which are not included in the simulation. A closer look at sequence dynamics underlying functional changes underscores the importance of indels: Over 40% of these events depend on indels in the sense that the energy change of the *cer* site caused by point substitutions alone would not have produced a loss of function. Moreover, the energy difference of the events with indels is on average larger than in the remainder of cases (Fig. 3 *E–G*). This is not surprising, because an indel can introduce a frameshift in the core motif and thus generate larger changes in energy and fitness than a single point mutation. This mode of evolution poses a challenge to modeling (which we have not attempted to address in this article) because neutral background rates of indels in regulatory regions are difficult to estimate.

Discussion

Binding Sites Evolve Under Epistasis. We have shown that the binding energy of a transcription factor target site, which is its most important biophysical variable, is also a key phenotypic measure of its evolution. Natural selection can be described as a *nonlinear* fitness landscape depending on the binding energy (Fig. 2*C*). This nonlinearity generates epistasis within binding sites as a natural consequence of the underlying biophysics: the regulatory effect of a site is a collective property of its sequence, and this functional coupling between nucleotides leads to their evolutionary coupling by selection. Salient features of promoter evolution, in particular, loss and gain of regulatory function by genetic drift, are consequences of this nonlinearity and cannot be captured by any *linear* fitness model. This is important to note because most computational methods for promoter analysis are based on position weight matrices. These methods assume explicitly, or more often implicitly, a linear evolution model leading to independent evolution of the nucleotides at different positions of a binding site. They are valuable for describing the sequence content of sites at the single-nucleotide level. However, quantifying the effects of selection on regulatory function benefits from identifying the phenotype on which selection acts most directly.

The reference to a specific molecular phenotype is what distinguishes this study from previous models of protein evolution, where epistasis between and within codons has been discussed (38–40) by using models where the selection effects of mutations are determined by rugged fitness landscapes (41), and the effects of epistasis are quantified by population genetic observables such as the index of dispersion. A particularly strong form of epistasis, where the fitness effect of a mutation changes its sign depending on the genetic context (42), has recently been discussed in the context of the evolution of antibiotic resistance (43). In comparison, our fitness landscape is a rather smooth function of the energy phenotype, as expected from the underlying biophysics. This makes the inference of epistasis a more subtle problem, which requires quantifying the phenotypic effect of individual mutations (and could not be addressed by using the index of dispersion). Moreover, our fitness landscape is a monotonic function of energy—a so-called mesa landscape (18)—which does not contain sign epistasis. Of course, the fitness landscape of an individual binding site contains sign epistasis if the site is tuned to weak binding; such fitness functions have been called crater landscapes (15). Averaging over many sites, however, wipes out any such signal, because most sites favor strong binding.

Epistatic interactions between binding sites have been addressed in a recent study of promoter evolution (44) where mutations within a given set of functional binding sites are taken to be neutral and a selection factor is associated to mutations in and out of these sets. In effect, the fitness landscape is taken to be a step function, an assumption on the form of selection that implies epistasis. In a broader context, epistasis is to be expected wherever selection acts on a quantitative phenotypic trait. Our study demonstrates this effect at the molecular level for TF binding sites: the method of selection inference used here should be applicable to other molecular phenotypes.

Genetic Drift or Adaptation? Our selection inference method assumes that observed energy differences between sites for the same factor, both within one species and between orthologous sites across species, are primarily due to genetic drift, rather than to differences in selection. The inferred fitness landscape describes moderate levels of selection, which allow for substantial phenotype divergence by genetic drift. This includes loss of binding function and leads to a finite functionality lifetime for promoters. The lifetime of promoters in yeast is approximately 10 times the inverse neutral point mutation rate.

Our model explains the observed phenotype evolution in a consistent way, but it is not guaranteed to be true. Alternatively, could many binding sites be fine-tuned to specific energy values, and fluctuations around these values be suppressed by strong selection? This hypothesis can be rejected for time-independent selection, because it cannot explain the substantial energy divergence between orthologous sites displayed in Fig. 3. However, strong site-dependent selection cannot be ruled out if it is also time-dependent, so that cross-species divergence between orthologous sites primarily reflects the adaptive response to changes in the fitness landscape and not genetic drift (this scenario has been assumed in the analysis of ref. 20). We note that evidence for selection fluctuations on evolutionary time scales has been established recently for coding and intergenic sequence in *Drosophila* (36). Disentangling the contributions of genetic drift and adaptation in promoter evolution is an outstanding challenge, which is important for quantifying the broader role of regulation in macroevolution. Here, we produce a quantitative estimate of the level of selection under a model of genetic drift, which can serve as benchmark for future studies of different models.

Loss and Gain of Function Involves Complex Sequence Dynamics. We have shown that sequence insertions and deletions play an important role in generating functional changes in promoters. For example, $>40\%$ of the observed loss-of-function events for regulation by Abf1 would not have occurred by point mutations alone. A single insertion or deletion can create or destroy even a strong binding site, an event that might have required several point mutations for its accomplishment. These events accelerate the relaxation to equilibrium in a stationary fitness landscape, as well as adaptive responses in a time-dependent landscape. The tempo of promoter evolution becomes even faster if we assume that a given promoter contains many equivalent ''shadow sites,'' any one of which can mutate into an actual binding site with the same regulatory function (15, 17). This mode of evolution leads to site turnover, where the binding function of a promoter is conserved between species but swapped between nonorthologous sites, and we do find occasional site turnover in yeast. However, regulatory sequence is notoriously ambiguous to align, and it is often difficult to decide whether a pair of sites is orthologous or not. The sequence analysis suffers from systematic errors, and the specific selection on binding sites needs to be taken into account (45). Our fitness function provides a specific scoring system for point mismatches and gaps within sites. Including these effects in a viable alignment method is a challenge for future work.

Methods

Alignment and Inference of Orthologous Sites. We obtain multiple alignments of intergenic sequences of the four yeast species *cer*, *par*, *mik*, and *bay* provided by ref. 21 by using the Prank alignment tool (46) [use of ClustalW (47) gives equivalent results]. To infer conserved binding sites, we scan the four-species alignment for low-energy sites ($E < \tilde{E} = 0.9$) in any one species and look for an orthologous low-energy site in any other species within ±10 base pairs from the mapped sequence (see ref. 20 for a similar approach). This procedure takes into account possible local alignment errors, which could produce false positive loss events (most conserved sites do not require this ''wobble''). To identify putative loss-of-function events, we look for orthologous promoters with a unique low-energy site present in three species, but no low-energy site present in the fourth species within an alignment window of size 140 [a typical promoter length in yeast (32)] around the mapped sequence. This criterion excludes cases where a promoter has more than one site.

Our inference procedures for the energy change $\Delta E = E' - E$ of a loss event and of the role of indels are detailed in *SI Text*.

Stochastic Modeling of Sequence Evolution. The neutral mutation process is inferred by recording a (4×4) substitution frequency matrix from aligned intergenic regions of *cer* and *bay*, which is translated into a mutation rate matrix $\mu_{\mathrm{a} \to \mathrm{b}}$ with different transition and transversion rates by using the HKY model (48). An alternative HKY model based on the sequence immediately flanking the $\mathcal{Q}_4^{\mathrm{no}}$ sites is discussed in *SI Text* (see also Fig. S6 *b* and *c*. The fitness landscape in the binding regime is inferred by Eq. **2** from the energy distributions $Q_{\mathrm{f}}(\mathrm{E})$ (obtained from the ensemble $\mathcal{Q}_4^{\mathrm{ol}}$) and $P_0(E)$ (obtained by generating random sequence with genomic single-nucleotide frequencies [0.320, 0.183, 0.163, 0.334]). These quantities, taken together, determine the Kimura–Ohta substitution rates within binding sites, $u_{\mathbf{a} \to \mathbf{b}} = \mu_{\mathbf{a} \to \mathbf{b}} 2N\Delta F_{\mathbf{ab}}/(1 - \exp[-2N\Delta F_{\mathbf{ab}}])$ (34). Beneficial changes ($\Delta F > 0$) have increased rates and deleterious ones ($\Delta F < 0$) have decreased rates. We simulate this process over evolutionary distances τ along the phylogenetic tree of ref. 21 (see Fig. 1) for site sequences with initial energy $E < \tilde{E}$ to obtain the predicted distributions $\Omega_f^{\tau}(\Delta E)$ and $\Omega_{\mathrm{loss}}^{\tau}(\Delta E)$ of the energy change $\Delta E = E' - E$ for conserved sites ($E' < \tilde{E}$) and loss events ($E' > \tilde{E}$), respectively. For details, see *SI Text* and Fig. S7.

ACKNOWLEDGMENTS. We thank the Kavli Institute for Theoretical Physics for their hospitality. This work was supported in part by Deutsche Forschungsgemeinschaft Grant SFB 680 (to V.M. and M.L.), National Institutes of Health Grant P50 GM071508 (to J.K. and C.C.), Department of Energy Grant DE-FG02-91ER40671 (to C.C.), by National Science Foundation Grant PHY-0650617 (to C.C.), and National Science Foundation Grant PHY05-51164 to the Kavli Institute for Theoretical Physics (University of California, Santa Barbara, CA).

1. Ptashne M, Gann A (2002) *Genes and Signals* (Cold Spring Harbor Lab. Press, Woodbury, NY).
2. Monod J, Jacob F (1961) Genetic regulatory mechanisms in the synthesis of proteins. *Cold Spring Harbor Symp Quant Biol* 26:389–401.
3. Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol Biol Evol* 12:1002–1011.
4. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403:564–567.
5. Wray GA, *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 20:1377–1419.
6. Ludwig MZ, *et al.* (2005) Functional evolution of a cis-regulatory module. *PLoS Biol* 3(4):e93.
7. Borneman AR, *et al.* (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317:815–819.
8. Fields D, He Y, Al-Uzri A, Stormo G (1997) Quantitative specificity of the Mnt repressor. *J Mol Biol* 271:178–194.
9. Stormo GD, Fields D (1998) Specificity, free energy and information content in protein DNA interactions. *Trends Biochem Sci* 23:109–113.
10. Lee TI, *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298:799–804.
11. Mukherjee S, *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 36:1331–1339.
12. Kinney JB, Tkačik G, Callan CG (2006) Precise physical models of protein–DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* 104:501–506.
13. Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22:141–149.
14. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233–237.
15. Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.
16. Mustonen V, Lässig M (2005) Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proc Natl Acad Sci USA* 102:15936–15941.
17. Lässig M (2007) From biophysics to evolutionary genetics: Statistical aspects of gene regulation. *BMC Bioinformatics* 8(Suppl 6):S7.
18. Gerland U, Hwa T (2002) On the selection and evolution of regulatory DNA motifs. *J Mol Evol* 55:386–400.
19. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol* 3(1):19.
20. Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3(5):0932–0942.
21. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.
22. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193:723–750.
23. Cho G, Kim J, Mo Rho H, Jung G (1995) Structure-function analysis of the DNA-binding domain of *Saccharomyces cerevisiae ABF1. Nucleic Acids Res* 23:2980–2987.
24. Wasserman W, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5:276–287.
25. McCue LA, *et al.* (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* 29:774–782.
26. Cliften P, *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76.
27. Harbison CT, *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.
28. Brown CT, Callan CG (2004) Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli. *Proc Natl Acad Sci USA* 101:2404–2408.
29. Doniger SW, Huh J, Fay JC (2005) Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res* 15:701–709.
30. MacIsaac KD, *et al.* (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* 7:113.
31. Pachkov M, Erb I, Molina M, van Nimwegen E (2007) SwissRegulon: A database of genome-wide annotations of regulatory sites. *Nucleic Acid Res* 35:D127–D131.
32. Erb I, van Nimwegen E (2006) Statistical features of yeast's transcriptional regulatory code. *IEEE Proceedings of the first International Conference on Computational Systems Biology (ICCSB)* 111–118.
33. Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61:763–771.
34. Kimura M, Ohta T (1971) Protein polymorphism as a phase of molecular evolution. *Nature* 229:467–469.
35. Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor Popul Biol* 10:254–275.
36. Mustonen V, Lässig M (2007) Adaptations to fluctuating selection in Drosophila. *Proc Natl Acad Sci USA* 104:2277–2282.
37. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol Biol Evol* 15:910–917.
38. Ohta T (1997) Role of random genetic drift in the evolution of interactive systems. *J Mol Evol* 44:S9–S14.
39. Ohta T (1998) Evolution by nearly-neutral mutations. *Genetica* 102/103:83–90.
40. Tachida H (2000) Molecular evolution in a multisite nearly neutral mutation model. *J Mol Evol* 50:69–81.
41. Kauffman S (1993) *The Origins of Order* (Oxford Univ Press, Oxford).
42. Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59:1165–1174.
43. Weinreich DM, Delaney NF, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114.
44. Raijman D, Shamir R, Tanay A (2008) Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comp Biol* 4(1):e7.
45. Sinha S, He X (2007) MORPH: Probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol* 3(11):e216.
46. Chenna R, *et al.* (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31:3497–3500.
47. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102:10557–10562.
48. Hasegawa M, Kishino H, Yano T-A (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.