Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies

Ville Mustonen and Michael Lässig*

Institut für Theoretische Physik, Universität zu Köln, Zülpicherstrasse 77, 50937 Cologne, Germany

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved August 8, 2005 (received for review June 30, 2005)

We study the evolution of transcription factor-binding sites in prokaryotes, using an empirically grounded model with point mutations and genetic drift. Selection acts on the site sequence via its binding affinity to the corresponding transcription factor. Calibrating the model with populations of functional binding sites, we verify this form of selection and show that typical sites are under substantial selection pressure for functionality: for cAMP response protein sites in *Escherichia coli*, the product of fitness difference and effective population size takes values $2N\Delta F$ of order 10. We apply this model to cross-species comparisons of binding sites in bacteria and obtain a prediction method for binding sites that uses evolutionary information in a quantitative way. At the same time, this method predicts the functional histories of orthologous sites in a phylogeny, evaluating the likelihood for conservation or loss or gain of function during evolution. We have performed, as an example, a cross-species analysis of *E. coli*, *Salmonella typhimurium*, and *Yersinia pseudotuberculosis*. Detailed lists of predicted sites and their functional phylogenies are available.

Regulatory interactions between genes are believed to provide an important mode of evolution, which accounts for a substantial part of the differentiation between species (1). This is reflected by the sequence variability of regulatory DNA: there is ample case evidence of compensatory evolution at conserved function but also of rapid functional changes even between closely related species (2). Lacking a quantitative model of regulatory evolution, however, alignments of regulatory sequences and predictions of their functionality have proven notoriously difficult.

A large body of existing work has focused on the identification of transcription factor-binding sites as the main functional elements of regulatory DNA. For factors with known binding specificity (given in the form of a position weight matrix), putative binding sites are identified from their conservation in cross-species comparisons. Different measures of conservation have been introduced, which involve, e.g., the sequence similarity of aligned loci or their independent high scoring in all species compared (3–7). These methods are powerful prediction tools for binding sites. From an evolutionary point of view, however, the conservation criteria are heuristic. Hence, it is difficult to quantify the statistical significance of the results, which depends on the number and evolutionary distance of the species compared. Sequence conservation tends to be too restrictive in cases where substantial sequence variation is compatible with the position weight matrix, in particular for distant species. Simple sequence similarity measures implicitly assume neutral evolution, whereas independent scoring of orthologous sites ignores the evolutionary link between the species altogether. Most importantly, none of these conservation measures allows a consistent statistical treatment of functional innovations in the evolution of binding sites.

A more explicit model for regulatory DNA should address two issues: How does the sequence divergence between species depend on their evolutionary distance, and how does the specific biological function of binding sites enter? Answering these questions is a considerable task for experiment and theory, which must link the biophysics of binding sites with their population dynamics. In particular, it involves quantifying the selection by which the sequence evolution at functional loci is distinguished from that of neutral background DNA. As an important step in this direction, the notion of a fitness landscape for binding-site sequences has been introduced, where the fitness of a site depends on the binding energy of the corresponding factor (8). The evolutionary importance of the binding energy has also been highlighted in ref. 9, where it was shown that nucleotide substitution rates within functional sites in *Escherichia coli* depend on the energy difference induced by the substitution as predicted from the position weight matrix. The biophysics of factor-DNA binding imposes stringent constraints on the form of the fitness landscape (10) and has important consequences for bioinformatic binding site searches (11). Using such fitness landscapes, we have introduced a stochastic evolution model for functional loci, which is based on Kimura–Ohta point substitutions with rates governed by the fitness difference between the corresponding sequence states (12, 13). This model demonstrates the possibility of rapid adaptive formation of binding sites under positive selection and provides evolutionary constraints on eukaryotic promoter architecture. A similar evolutionary model (14) underlies a recently introduced method to identify conserved binding sites in multiple alignments (15).

In this paper, we develop a quantitative evolutionary rationale for the cross-species analysis of regulatory sequences, which goes beyond the mere prediction of binding sites. For aligned regulatory DNA of orthologous genes, our method predicts sites together with their functional evolution. The method is based on the evolution model of refs. 12 and 13 and uses a bioinformatic measurement of selection pressures for functionality, which is obtained from sequence data of verified functional sites. Typical functional loci for pleiotropic factors, as exemplified by the cAMP response protein (CRP) family in *E. coli*, are found to be under substantial selection, in contrast to nonfunctional loci, which evolve neutrally. For families of aligned loci, our method assigns likelihood values to different modes of evolution and associates them with functional histories: (*i*) neutral evolution of nonfunctional loci, (*ii*) evolution of functional loci under time-independent selection, and (*iii*) evolution under time-dependent selection, corresponding to loss or gain of function along a given branch of the phylogeny.

Theory

Evolution Models for Nonfunctional Sequence and Functional Loci. We consider genomic loci $\mathbf{a} = (a_1, \ldots, a_l)$ consisting of l contiguous nucleotides and elementary substitution processes $\mathbf{a} \to \mathbf{b}$, where $\mathbf{a}$, $\mathbf{b}$ are any two sequence states differing by exactly one nucleotide. For nonfunctional (background) sequences, we use uniform nucleotide substitution rates $\mu_{\mathbf{a}\to\mathbf{b}}$ depending on the nucleotide to be mutated and on its nearest sequence neighbors (16). Models of this type are neutral with respect to factor binding and have been shown to provide a good description of intergenic background DNA in *E. coli* (11). A locus is defined as functional if binding of the corresponding factor at that locus affects the regulation of a gene. Functional loci are assumed to be under selection. This is described

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: CRP, cAMP response protein.

*To whom correspondence should be addressed. E-mail: lassig@thp.uni-koeln.de.

www.pnas.org/cgi/doi/10.1073/pnas.0505537102

by a (Malthusian) fitness function $F(\mathbf{a})$, which measures the contribution of a genotype $\mathbf{a}$ to the growth rate of the number of individuals carrying that genotype (and is therefore defined only up to an additive constant, the genotype-independent fitness). Notice that this definition of a functional locus is weaker than that of a functional binding site, which is a functional locus with a sequence state $\mathbf{a}$ that is likely to actually bind the factor. A functional locus can lose its binding sequence due to deleterious mutations, and conversely, a nonfunctional locus can become a spurious binding site. According to the Kimura–Ohta theory (17–19), selection leads to modified substitution rates at functional loci,

$$u_{\mathbf{a}\to\mathbf{b}} = \mu_{\mathbf{a}\to\mathbf{b}} N \frac{1 - \exp[-2(F(\mathbf{b}) - F(\mathbf{a}))]}{1 - \exp[-2N(F(\mathbf{b}) - F(\mathbf{a}))]}, \qquad [1]$$

where N denotes the effective population size. In writing Eq. **1**, we have assumed $\mu N \ll 1$, so that subsequent substitution processes are well separated in time and can be assumed to be independent.

Stationary Population Distributions and Evolutionary Scoring. For background sequences, we use a stationary distribution of the form (11)

$$P_0(\mathbf{a}) = p_0(a_1) \prod_{i=2}^{l} \pi_0(a_i|a_{i-1}), \qquad [2]$$

where $p_0(a)$ is the single-letter equilibrium distribution, and $\pi_0(a|a')$ is the conditional distribution for letter a given its left neighbor a' (see *Supporting Text*, which is published as supporting information on the PNAS web site). Assuming the underlying neutral dynamics with rates $\mu_{\mathbf{a}\to\mathbf{b}}$ satisfies detailed balance, the dynamics under selection satisfies detailed balance as well, and the stationary distribution for functional loci takes the form (12, 13)

$$Q(\mathbf{a}) = P_0(\mathbf{a}) \exp[2NF(\mathbf{a}) + \text{const.}], \qquad [3]$$

with the constant given by the normalization $\sum_{\mathbf{a}} P_0(\mathbf{a}) = \sum_{\mathbf{a}} Q(\mathbf{a}) = 1$. These distributions give the probability density to find a locus with sequence $\mathbf{a}$, which can be inferred from long-term frequency counts at a given locus, or equivalently, from the fraction count of sequence $\mathbf{a}$ in a large ensemble of independently evolving loci. We identify the distributions P_0 and Q with the ensembles of background respective functional loci for a given factor in a genome. (This involves the approximation that all functional loci are under similar selection pressure.) Hence, the usual position weight matrix (20) for the factor is given by the single-nucleotide (marginal) distributions

$$q_i(a) = \sum_{a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_l} Q(\mathbf{a}). \qquad [4]$$

If the background distribution is approximated by a factorized form, $P_0(\mathbf{a}) = \Pi_{i=1}^{l} p_0(a_i)$, and if the fitness is an additive function of the nucleotide positions, $F(\mathbf{a}) = \sum_{i=1}^{l} f_i(a_i)$, the Q distribution factorizes as well, $Q(\mathbf{a}) = \Pi_{i=1}^{l} q_i(a_i)$ with $q_i(a) = p_0(a) \exp[2N f_i(a) + \text{const.}]$. This form of the stationary distribution has previously been introduced in ref. 14 for protein-coding sequences and has been used in ref. 15. For binding sites, however, the factorized form is a heuristic approximation, because generic fitness landscapes are not additive. Regulatory fitness effects follow from the expression level of the regulated gene, which in turn depends on the relevant binding sites through the binding probability of the corresponding transcription factor (8, 13). The factor-binding energy is often nearly additive in the nucleotide positions (21). The individual contributions $\varepsilon_i(a)$ can be inferred in an approximate way from the position weight matrix (20, 22) (up to an overall constant ε_0),

$$E = \sum_{i=1}^{l} \varepsilon_i(a_i) \quad \text{with } \varepsilon_i(a) = \varepsilon_0 \log \frac{q_i(a)}{p_0(a)}. \qquad [5]$$

The binding probability, however, is a strongly nonlinear function of the energy. Hence, the fitness effect of a substitution at one position depends on the nucleotides present at all other positions. This induces correlations between nucleotide frequencies at any two positions within functional loci (13), in addition to the short-ranged correlations already present in background sequences. These correlations prevent the factorization of the distributions $P_0(\mathbf{a})$ and $Q(\mathbf{a})$. However, because the fitness $F(\mathbf{a})$ depends on the sequence state $\mathbf{a}$ only via the binding energy $E(\mathbf{a})$, we can project these distributions on the energy E as independent continuous variables, summing over all sequence states $\mathbf{a}$ with an (approximately) equal value of E. Denoting the projected ensembles for simplicity with the same symbols P_0 and Q, Eq. **3** takes the form

$$Q(E) = P_0(E) \exp[2NF(E) + \text{const.}]. \qquad [6]$$

It is this simplification that enables us to infer the functional form of these distributions from bioinformatic frequency counts. The total distribution of energies in the noncoding part of the genome is

$$W(E) = (1 - \lambda)P_0(E) + \lambda Q(E). \qquad [7]$$

From a bioinformatic point of view, this is a hidden Markov model for the sequence composition of noncoding DNA. The two alternative distributions $P_0(E)$ and $Q(E)$ have prior probabilities $1 - \lambda$ and λ, i.e., the parameter λ measures the overall fraction of the genome covered by functional loci. The relative likelihood between the distributions Q and P_0 is described by the score function $S(E) \equiv \log[Q(E)/P_0(E)]$. Comparing with Eq. **6**, we obtain the identification

$$S(E) = 2NF(E) + \text{const.}, \qquad [8]$$

which establishes an important and rather general evolutionary grounding of the bioinformatic log-likelihood score.

Time-Dependent Distributions and Cross-Species Scores. It is straightforward to generalize the probabilistic analysis to pairs of species separated by an evolutionary time t. Defining the conditional transition probabilities $G_0^t(E_2|E_1)$ and $G_s^t(E_2|E_1)$, we obtain the joint distribution of energy pairs for orthologous loci

$$P_0^t(E_1, E_2) = G_0^t(E_2|E_1)P_0(E_1), \qquad [9]$$

at neutrality and

$$Q^t(E_1, E_2) = G_s^t(E_2|E_1)Q(E_1), \qquad [10]$$

under constant selection. The total distribution of energy pairs is

$$W^t(E_1, E_2) = (1 - \lambda)P_0^t(E_1, E_2) + \lambda Q^t(E_1, E_2). \qquad [11]$$

The ensembles P_0^t and Q^t define the log-likelihood score

$$S^t(E_1, E_2) = \log \frac{Q(E_1)}{P_0(E_1)} + \log \frac{G_s^t(E_2|E_1)}{G_0^t(E_2|E_1)}, \qquad [12]$$

which is easily shown to be symmetric in its two arguments, because detailed balance ensures reversibility of the evolutionary dynamics at equilibrium. Eq. **12** is a key result of this paper: it shows how the specific evolution of binding loci can be integrated into a bioinformatic scoring procedure. The evolutionary information is contained in the second score term, i.e., the log ratio of the transition probabilities. This term measures a difference in evolutionary fates: A locus with energy E_1 in the "twilight region" between the Q and P_0 ensembles on average maintains its binding energy if functional but evolves toward lesser binding if nonfunctional. The energy transition probabil-

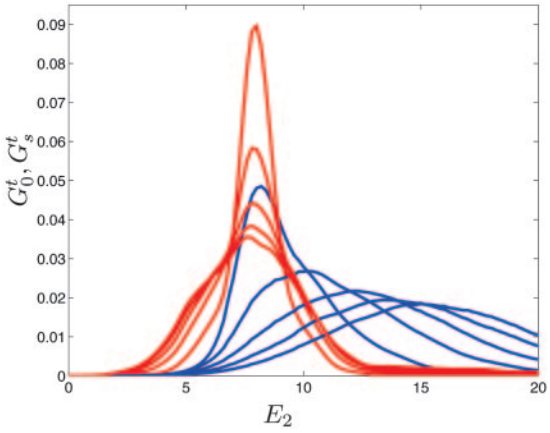

Fig. 1. Cross-species energy transition probabilities $G_0^t(E_2|E_1)$ for neutral evolution (blue) and $G_s^t(E_2|E_1)$ for evolution under time-independent selection in the fitness landscape of Fig. 3*a* (red). Curves are shown for fixed initial energy $E_1 = 8$ and various evolutionary distances t. The third curve in each family belongs to the distance between aligned loci of *E. coli* and *S. typhimurium*. Typical loci evolve toward weaker binding under neutrality but maintain their binding energy under selection.

ities $G_0^t(E_2|E_1)$ and $G_s^t(E_2|E_1)$ can be obtained in a straightforward way from the underlying sequence evolution process (see *Supporting Text*). They are shown in Fig. 1 for fixed E_1 and for various evolutionary distances t. The difference between the distributions G_0^t and G_s^t, i.e., the additional discriminatory power of the evolutionary information, is seen to increase with the distance of the species compared. In the long-distance limit, we have $G_0^t(E_2|E_1) \to P_0(E_2)$ and $G_s^t(E_2|E_1) \to Q(E_2)$, leading to independent scoring of aligned loci in Eq. **12**.

Time-Dependent Selection and Functional Switching. Here we generalize the evolutionary model to include loss or gain of function at the level of individual loci. Consider a rooted phylogeny consisting of two species at evolutionary distances t_1 and t_2 from their last common ancestor, i.e., at distance $t = t_1 + t_2$ from each other. We assume that an initially functional locus can lose function at a small rate ν_- (with $\nu_- t \ll 1$), and conversely, an initially nonfunctional locus can gain function at a comparable rate ν_+ (such that the average fraction λ of functional loci is maintained). There are now four alternative evolutionary histories involving at most one functional switch: evolution under time-independent neutrality or selection, time-dependent selection leading to functionality at t_1, and nonfunctionality at t_2, and vice versa (see Fig. 2). These occur with probabilities $\lambda_0^{\mathbf{t}} = (1-\lambda)(1-\nu_+ t)$, $\lambda_Q^{\mathbf{t}} = \lambda(1-\nu_- t)$, $\lambda_{s0}^{\mathbf{t}} = (1-\lambda)\nu_+ t_1 + \lambda\nu_- t_2$, and $\lambda_{0s}^{\mathbf{t}} = (1-\lambda)\nu_+ t_2 + \lambda\nu_- t_1$, respectively, with the abbreviation $\mathbf{t} = (t_1, t_2)$. The four corresponding energy pair distributions are $P_0^t(E_1, E_2)$, $Q^t(E_1, E_2)$,

$$R_{s0}^{\mathbf{t}}(E_1, E_2) = \frac{1}{\lambda_{s0}^{\mathbf{t}}} \int dE' \times \left[(1-\lambda)\nu_+ \int_0^{t_1} dt' G_s^{t_1-t'}(E_1|E') G_0^{t_2+t'}(E'|E_2) P_0(E_2) + \lambda\nu_- \int_0^{t_2} dt' G_0^{t_2-t'}(E_2|E') G_s^{t_1+t'}(E'|E_1) Q(E_1) \right], \quad [13]$$

and $P_{0s}^{\mathbf{t}}(E_1, E_2)$, which is defined in an analogous way. In Eq. **13** t' denotes the switching point and E' the energy at that point. We

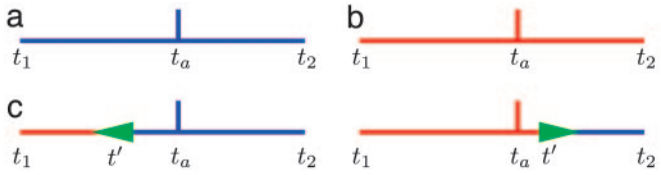

Fig. 2. Functional phylogenies for two species at evolutionary distances t_1 and t_2, counted from their last common ancestor at time $t_a = 0$. Branch segments with neutral evolution are shown in blue with evolution under selection in red. (*a*) Neutral evolution of nonfunctional loci, described by the energy pair distribution P_0^t. (*b*) Evolution of functional loci under time-independent selection, described by the distribution Q^t. (*c*) Evolution under time-dependent selection generating a functional locus in species 1 and a nonfunctional locus in species 2, described by the distribution $R_{s0}^{\mathbf{t}}$. This mode involves either a gain of function between ancestor and 1 or a loss of function between ancestor and 2. The switching event at time t' is denoted by a green arrow. The corresponding mode where the roles of the two species are interchanged is described by the distribution $R_{0s}^{\mathbf{t}}$.

have approximated the ancestral energy distribution by the stationary ensembles P_0 or Q and have used detailed balance. Within this switch mode, we have summed over the probabilities of gain and loss of function. Disentangling these alternatives is possible but statistically insignificant in phylogenies with few species. Neglecting histories with more than one functional switch, the total distribution of energy pairs is now

$$W^{\mathbf{t}}(E_1, E_2) = \lambda_0^{\mathbf{t}} P_0^t(E_1, E_2) + \lambda_Q^{\mathbf{t}} Q^t(E_1, E_2) + \lambda_{s0}^{\mathbf{t}} R_{s0}^{\mathbf{t}}(E_1, E_2) + \lambda_{0s}^{\mathbf{t}} R_{0s}^{\mathbf{t}}(E_1, E_2), \quad [14]$$

and there are three independent log-likelihood scores weighing each of the ensembles $Q^t, R_{s0}^{\mathbf{t}}, R_{0s}^{\mathbf{t}}$ against the background ensemble P_0^t. The hidden Markov model can readily be extended to rooted phylogenies with more than two species. The expressions for the P_0, Q, and R distributions generalizing Eqs. **9**, **10**, and **13** involve a factor G_0 for each branch under neutral evolution and a factor G_s for each branch under selection, as well as integrations over the (unobserved) energies at the internal nodes of the phylogeny and the variables E' and t' of the switching point. In the case of three species, there are eight different functional histories: time-independent neutrality and selection, as well as time-dependent selection leading to a functional locus in any one or any two species (see *Supporting Text* for details).

Site Prediction and Quality Measures. For a given pair of aligned loci with energies (E_1, E_2), the hidden Markov model (Eq. **14**) determines the probability of belonging to each of its four ensembles. The probability of conserved functionality is

$$\rho_Q^{\mathbf{t}}(E_1, E_2) = \frac{\lambda_Q^{\mathbf{t}} Q^t(E_1, E_2)}{W^{\mathbf{t}}(E_1, E_2)}. \quad [15]$$

The corresponding probabilities $\rho_0^{\mathbf{t}}(E_1, E_2)$ for conserved neutrality, $\rho_{s0}^{\mathbf{t}}(E_1, E_2)$ and $\rho_{0s}^{\mathbf{t}}(E_1, E_2)$ for functional switching, and $\rho_Q(E)$ for functionality in the single-species case are defined in an analogous way. These probabilities form the basis of our predictions for individual sites and site pairs, and they serve to quantify the predictive quality. For functional loci, we further distinguish functional and nonfunctional binding sites, using as approximate threshold the energy $E_{\max}$ of the weakest verified site. Hence, given a total of n aligned pairs of loci, an expected number

$$n_f^{\mathrm{tot}} = n \int_{\rho_Q^{\mathbf{t}} > \rho_{\min}} dE_1 dE_2 \, \rho_Q^{\mathbf{t}}(E_1, E_2) W^{\mathbf{t}}(E_1, E_2) \quad [16]$$

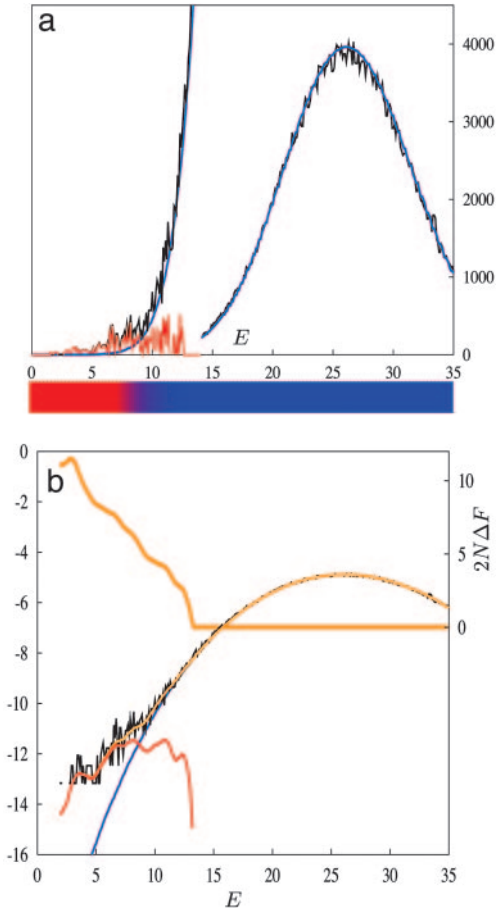

Fig. 3. Energy statistics and fitness landscape for CRP-binding loci in *E. coli*. (*a*) Count histogram with energy bins of width 0.1 (black), expected background counts (blue), and excess counts above background (red), with a 30-fold zoom into the region $E < 14$. The color bar indicates the probability of functionality $\rho_Q(E)$, ranging from 1 (red) to 0 (blue). (*b*) Decomposition of the counts (log-scale, left *y* axis) according to the single-species hidden Markov model: background distribution $(1 - \lambda)P_0(E)$ (blue), distribution $\lambda Q(E)$ of functional loci (red), and total distribution $W(E)$ (orange). The resulting fitness landscape $\Delta F(E)$ according to Eq. **6** is also shown in orange (thick curve, right *y* axis).

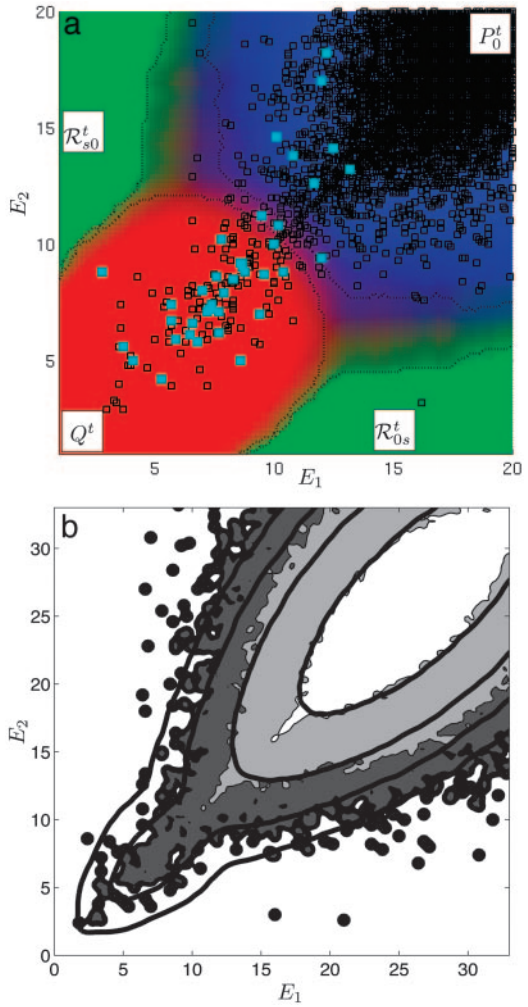

Fig. 4. Binding energy pairs and functional histories for aligned loci in *E. coli* and *S. typhimurium*. (*a*) Dot plot of counts (E_1, E_2), including verified binding sites (light blue). The background color shading indicates the likelihood of functional histories, varying between blue (conserved neutrality), red (conserved function), and green (functional switching). Isoprobability lines $\rho_\alpha^t = 0.55$ ($\alpha = 0, Q, 0s, s0$) are dotted. (*b*) Energy pair density obtained from the counts (filled contours), compared with the distribution $W^t(E_1, E_2)$ (contour lines, $W^t = 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$).

with $\rho_{\min} = \rho_Q^t(E_{\max}, E_{\max})$ are conserved functional sites. (Because ρ_Q^t is small for entries of order $E_{\max}$, this number depends only weakly on the cutoff $\rho_{\min}$.) A predicted set of functional site pairs with $\rho_Q^t(E_1, E_2) > \rho_0 \geq \rho_{\min}$ contains

$$n(\rho_0) = n \int_{\rho_Q^t > \rho_0} dE_1 dE_2 W^t(E_1, E_2) \qquad [17]$$

entries, of which $n_f(\rho_0)$ are expected to be true functional pairs. The number $n_f(\rho_0)$ is given by the integral of Eq. **16** over the region $\rho_Q^t > \rho_0$. Hence, the expected fractions of false positives and of false negatives are

$$\gamma_+(\rho_0) = 1 - \frac{n_f(\rho_0)}{n(\rho_0)}, \quad \gamma_-(\rho_0) = 1 - \frac{n_f(\rho_0)}{n_f^{\text{tot}}}. \qquad [18]$$

Analogous definitions apply to the single-species case.

Results

Selection Pressure for CRP Sites in *E. coli*. Scanning the genome of *E. coli* (obtained from the NCBI database, accession no. NC_000913) produces sequence counts of $n = 520{,}729$ loci in 4,244 intergenic regions. We use a relatively large window size of $l = 22$, taking into account core binding motifs as well as informative flanking positions. Our CRP position weight matrix $q_i(a)$ ($i = 1, \ldots, l$; $a = A, C, G, T$) is obtained from 48 experimentally verified binding sites in the DPInteract database (23). For each sequence count $\mathbf{a} = (a_1, \ldots, a_l)$, we hence infer the CRP-binding energy $E(\mathbf{a})$ from Eq. **5** (in units of ε_0 and with $E = 0$ set to the point of maximal binding). The resulting energy histogram is shown in Fig. 3*a*. In the region $E > E_{\max} \approx 13$, where no factor binding is expected, the data are well approximated by the background distribution $P_0(E)$, whereas the excess counts for $E < E_{\max}$ are attributed to functional loci. Their distribution $Q(E)$ in this region can be estimated independently from the ensemble of verified binding sites. Optimal consistency with the hidden Markov model (Eq. **7**) is reached for $\lambda = 0.00065$, where the distribution $W(E)$ produces a very similar form of $Q(E)$ and fits the entire histogram well; see Fig. 3*b*. The effective fitness landscape for functional loci is then inferred from the data using Eq. **6**. In the nonbinding region $E > E_{\max}$, the fitness takes a constant value F_0, i.e., the evolution is always neutral in that region, as expected. The constant F_0 is unimportant, because only fitness

differences enter the evolutionary dynamics (Eq. **1**). Larger values occur in the binding region $E < E_{\max}$. The excess fitness landscape $2N\Delta F(E) \equiv 2N(F(E) - F_0)$ is shown as the orange line in Fig. 3*b*. Loci with strong binding (i.e., with energies $E < 8$) have substantial effective fitness values in the range $2N\Delta F \sim 6$–11, which are interpreted as typical selection pressures for functionality. Genetic drift counteracts selection, producing also loci with weaker binding ($8 < E < 12$) and reduced effective fitness $2N\Delta F < 6$. These fitness estimates are rather robust results of our procedure. It should be kept in mind, of course, that the shape of the fitness landscape $\Delta F(E)$ reflects an average over the family of CRP-binding sites, which have a spectrum of individual selection coefficients and selected binding strengths.

With this fitness landscape, the hidden Markov model (Eq. **7**) thus gives an excellent description of the CRP-binding energy statistics in intergenic DNA of *E. coli*. For an individual locus, the model predicts the probability $\rho_Q(E)$ of being functional, given its binding energy E. This probability is indicated by the color shading in Fig. 3*a*. The parameter λ determines the total number of predicted functional binding sites, $n_f^{\rm tot} \sim 340$. The power to predict individual binding sites remains limited, however. The reason is apparent from Fig. 3*a*: Many functional sites have energy values in the "twilight region" (appearing in violet), where there is already a sizeable amount of background counts. Hence, any prediction will be torn between many false negatives or many false positives, depending on the energy cutoff chosen. This situation will be drastically improved by the cross-species analysis to which we now turn.

Evolution Between *E. coli* and *Salmonella typhimurium*. The *Salmonella* genome is also obtained from NCBI (accession no. NC_003197). Our alignment of the two genomes contains 135,534 pairs of loci in well aligned intergenic regions flanked by orthologous genes (for details, see *Supporting Text* and refs. 24 and 25). The average identity between aligned sequences is 93%, which measures the evolutionary distance t between the two species. The CRP-binding energies E_1 in *E. coli* and E_2 in *S. typhimurium* are inferred using the same position weight matrix, which is justified, because the factor itself is highly conserved (3, 9). The resulting dot plot of energy pairs (E_1, E_2) is shown in Fig. 4*a*. The distribution is significantly pinched to the diagonal in the binding region $E_1, E_2 < 12$, indicating the expected higher conservation of the energy for functional binding sites (9). We quantify this effect by the conditional probability $G_s^t(E_2|E_1)$ of evolving from energy E_1 to E_2 under selection as compared with its neutral counterpart $G_0^t(E_2|E_1)$ (see *Theory*). Both distributions are readily obtained from numerical simulations of the evolution processes; examples are shown in Fig. 1.

Fig. 4*b* contains again the energy pair counts together with the distribution $W^{\rm t}(E_1, E_2)$ of the two-species hidden Markov model (Eq. **14**). We use the same fitness landscape $F(E)$ as for *E. coli* and a midpoint approximation for the root point of the tree, i.e., $t_1 = t_2 = t/2$. The fit parameter $\lambda = 0.0018$ is now higher than in the single-species case (reflecting a higher fraction of functional loci in aligned regions), and there are small probabilities of selection gain and loss, $\nu_+ t \sim \nu_- t \sim 0.025$. Quite remarkably, this distribution reproduces the entire energy pair data well, which indicates the consistency of our approach as well as an overall similarity of the evolutionary characteristics between the two species compared.

Functional Histories. For an individual pair of aligned loci with energies (E_1, E_2), the hidden Markov model (Eq. **14**) predicts the probabilities of conserved neutrality and conserved function, $\rho_0^{\rm t}$ and $\rho_Q^{\rm t}$, and of functional switching, $\rho_{s0}^{\rm t}$ and $\rho_{0s}^{\rm t}$; see Eq. **15**. These probabilities are indicated by the color shading in Fig. 4*a*. The twilight region between the Q^t and P_0^t ensembles (appearing in violet) is seen to be much smaller and pushed toward larger energies $(E_1, E_2 \sim 10)$ than in Fig. 3*a*. This indicates that the additional evolutionary information substantially improves the predictions already for two species. For example, a prediction list for conserved functionality is obtained by ranking the pairs of loci in order of decreasing $\rho_Q^{\rm t}$ with a lower cutoff ρ_0. It has estimated fractions $\gamma_+(\rho_0)$ of false positives and $\gamma_-(\rho_0)$ of false negatives, which are given by Eq. **18**. Plotting the two parameters against each other produces a so-called detection error trade-off curve (26), which can be compared with its single-species counterpart (see also ref. 11). Both curves are shown in Fig. 5, which is published as supporting information on the PNAS web site, which quantifies the predictive power gained by the cross-species comparison (see *Supporting Text*). The Q^t list with a cutoff $\rho_0 = 0.30$, which contains 211 pairs of loci, is available upon request (see *Supporting Text*). Among them are 32 of the 40 verified binding sites in the aligned region. Compared with recently published lists of sites in *E. coli* (9, 11) and to our own single-species analysis, this list is considerably shorter at a comparable detection level of true sites. Yet it contains a substantial number of new entries, which are statistically significant at the level of two species but not of a single one.

A fraction of the functional binding sites in one species is predicted to lose their binding ability during evolution due to deleterious mutations, although their loci remain functional, i.e., under conserved selection. From *E. coli* to *Salmonella* or vice versa, we estimate this fraction to be ≈5%, by using the energy transition probabilities $G_s^t(E_2|E_1)$. However, our model also contains another mode of functional switching, which is due to loss or gain of selection for a locus. The corresponding site pairs have widely differing energies (E_1, E_2), which lead to high values of $\rho_{0s}^{\rm t}$ or $\rho_{s0}^{\rm t}$. Examples include CRP loci in the intergenic regions *prmA-yhdG* and *ytfJ-ytfK*, which are predicted to be functional *E. coli* but not in *S. typhimurium*. In the first case, there is no other likely CRP-binding site in the same intergenic region, indicating a possible functional change in the promoter as a whole. The second case, where one finds also a conserved site in the same region, may point to a compensatory change between loci, which leaves the function of the promoter as a whole intact. The evolutionary analysis gets simpler if experimental information is available even in one of the species compared. For example, the second CRP-binding locus from DPInteract (23) for *nupG* has energies $E_1 = 2.8$ in *E. coli* and $E_2 = 8.8$ in *S. typhimurium*. Because we know that the site is functional in *E. coli*, we need to compare only the probabilities of evolution under constant and time-dependent selection, leading to a substantial likelihood for a functional switch. There is a conserved site in the same intergenic region, which could take over the binding function in *S. typhimurium*. Note that the statistical weight in favor of a switch stems solely from the large energy change; both sites would individually qualify as functional under standard independent scoring. Of course, these predictions bear a higher level of uncertainty, because alignment ambiguities can lead to an artificially high value of the energy difference, and the prior switching probabilities ν_+, ν_- are only order-of-magnitude estimates.

We have extended our analysis to include a third species, *Yersinia pseudotuberculosis* (NCBI accession no. NC_006155). Dot plots of energies (E_1, E_2, E_3) for triplets of aligned loci and their probabilistic scoring are reported in *Supporting Text* and Figs. 6 and 7, which are published as supporting information on the PNAS web site. As expected, we find a further improvement of the detection error tradeoff for prediction of loci with conserved function; see Fig. 5. This is due in part to the alignment, which introduces a bias toward conserved loci. We also find candidate loci with loss or gain of function, such as the fourth *malE-malK* locus in *E. coli*, which we predict to be nonfunctional in both *S. typhimurium* and in *Y. pseudotuberculosis*. Three other verified sites in that region are conserved in all three species. Similar candidates for functional switches are the second verified binding sites for *dadAX*, *tsx*, and *araB*.

Discussion

Binding Sites in Bacteria Evolve Under Substantial Selection. Our evolutionary model is based on stochastic substitutions in the space of sequences $(a_1, \ldots, a_l)$ of binding loci. These loci are treated as coherent population genetic units, taking into account that the evolution of any two nucleotides within a functional locus is correlated (13). This differs from standard bioinformatic approaches such as position weight matrices, which assume the nucleotides a_i to be independent. Our *in silico* measurement of the selection pressure is based on the sequence ensemble Q of functional loci and its background counterpart P_0. Assuming that functional loci evolve under selection, and background loci evolve neutrally, the log-likelihood score of these ensembles determines the effective fitness difference of sequence states at these two kinds of loci: $S \equiv \log(Q/P_0) = 2N\Delta F$, where N is the effective population size. For CRP loci in *E. coli*, we obtain effective fitness differences $2N\Delta F$ of order 10 between strong factor binding and no binding. Because our method involves ensembles, this is an order-of-magnitude estimate for typical loci, which does not exclude that some sites may be under substantially stronger or weaker selection. We note, however, that our estimate also predicts the correct amount of energy conservation for functional loci found in our cross-species analysis.

A substantial level of selection explains well known evolutionary characteristics of regulatory sequences (2): they may be well conserved between distant species (if under constant selection for functionality) but can also show considerable variation even between closely related species (if under positive selection for change). At the level of selection found, binding-site gain by rapid adaptive evolution as discussed in ref. 13 is indeed possible. On the other hand, conservation will not be complete even under selection. A certain fraction (increasing with evolutionary distance) of initially functional sites will be lost because of deleterious mutations. This opens the possibility of compensatory changes involving different loci, as they are observed in ref. 27. It also indicates that the theory of promoter evolution should not stop at the level of individual binding sites. Selection couples not only the nucleotides within one locus but also the evolutionary fate of different loci. Understanding the long-term dynamics of regulation ultimately requires a consistent population-genetic theory of entire promoters.

Improving Binding Site Searches Requires a Quantitative Evolutionary Rationale. The difficulty of predicting functional sites from their binding score in a single species is well known and has been called the "futility theorem" in ref. 28. It is caused by the coexistence of functional and nonfunctional loci in the twilight region of marginal binding. In the framework of our probabilistic model, this is quantified by tradeoff curves between false positives and false negatives. What is a computational dilemma, may, however, reflect evolutionary design. If a sufficient reservoir of marginal binding seeds is present even in background sequences, a fully functional site can form by rapid adaptation as a response to new demand, ensuring the evolvability of regulatory interactions (13).

To overcome the futility theorem, we have introduced here a quantitative method that includes evolutionary information into binding-site searches. At the core of our model are the cross-species energy transition probabilities G_s^t and G_0^t, which quantify the "phenotypic" evolution of loci and discriminate efficiently between functionality and nonfunctionality (see Fig. 1). These probabilities can be used to build a systematic likelihood score for families of aligned orthologous loci in a phylogeny, which is of the general form $S = \log(Q/P_0) + \log(G_s^t/G_0^t)$. This scoring allows clear significance estimates and rankings of the results.

We have applied the method to comparative analysis of three bacterial species. We find a substantial improvement of the predictive quality already at the level of two species and a further improvement for three species. This confirms the results of a recent study of conserved sites in several *Saccharomyces* species, where the significance of the evolutionary information as a function of evolutionary distances is discussed in detail (15). Of course, elementary evolutionary steps other than point mutations are expected to become important in eukaryotes. Nevertheless, our general rationale of inferring selection pressures from site frequencies should remain applicable.

Putative Regulatory Innovations in Bacterial Phylogenies Can Be Traced by Comparative Sequence Analysis. Previous approaches have focused on the conservation of regulatory sequences as a sign of their functionality. Here we aim at a more comprehensive view, which includes functional changes into a quantitative statistical model. We emphasize again that in the presence of selection, there are two conceptually different modes of change: binding sites can lose or gain functionality due to deleterious or beneficial mutations at constant selection for binding, or they can respond to changes in the selection itself. Our model distinguishes these modes statistically by their energy transition probabilities and thus builds functional phylogenies for specific loci (as exemplified in Fig. 2).

In our comparative analysis of bacterial species, we find a large number of loci predicted to have conserved function but also some cases with evidence for gain or loss of function. It has been shown that changes in the gene regulation of orthologous genes can lead to phenotypic differences between *E. coli* and *S. typhimurium* (29). With caveats due to uncertainties in the rates of loss or gain of function, our findings provide at least a starting point for further targeted cross-species experiments. We can thus begin to quantify the role of regulatory innovations in molecular evolution.

We thank Johannes Berg and Ulrich Gerland for a critical reading of the manuscript. V.M. acknowledges support through the STIPCO European network, Contract HPRN-CT-2002-00319 (grant to M.L.).

1. Ptashne, M. & Gann, A. (2002) *Genes and Signals* (Cold Spring Harbor Lab. Press, Woodbury, NY).
2. Wray, G. A., Hahn, M. W., Abouheif, H., Balhoff, J. P., Pizer, M., Rockman, M. V. & Romano, R. A. (2003) *Mol. Biol. Evol.* **20,** 1377–1419.
3. Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002) *Genet. Res.* **12,** 298–308.
4. McCue, L. A., Thompson, W., Carmack, C. S. & Lawrence, C. E. (2002) *Genet. Res.* **12,** 1523–1532.
5. Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N. & Wasserman, W. W. (2003) *J. Biol.* **2,** 13.
6. Clifton, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) *Science* **301,** 71–76.
7. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) *Nature* **423,** 241–254.
8. Gerland, U. & Hwa, T. (2002) *J. Mol. Evol.* **55,** 386–400.
9. Brown, C. T. & Callan, C. G., Jr. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2404–2409.
10. Gerland, U., Moroz, D. & Hwa, T. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 12015–12020.
11. Djordjevic, M., Sengupta, A. M. & Shraiman, B. I. (2003) *Genome Res.* **13,** 2381–2390.
12. Berg, J. & Lässig, M. (2003) *Biophysics (Moscow)* **48,** Suppl. 1, 36–44.
13. Berg, J., Willmann, S. & Lässig, M. (2004) *BMC Evol. Biol.* **4,** 42.
14. Halpern, A. L. & Bruno, W. J. (1998) *Mol. Biol. Evol.* **15,** 910–917.
15. Moses, A. M., Chiang, D. Y., Pollard, A. P., Iyer, N. I. & Eisen, M. B. (2004) *Genome Biol.* **5,** R98.
16. Arndt, P. & Hwa, T. (2005) *Bioinformatics* **21,** 2322–2328.
17. Kimura, M. (1962) *Genetics* **47,** 713–719.
18. Kimura, M. & Ohta, T. (1969) *Genetics* **61,** 763–771.
19. Ohta, T. & Tachida, H. (1990) *Genetics* **126,** 219–229.
20. Berg, O. & von Hippel, P. (1987) *J. Mol. Biol.* **193,** 723–750.
21. Fields, D., He, Y., Al-Uzri, A. & Stormo, G. (1997) *J. Mol. Biol.* **271,** 178–194.
22. Stormo, G. D. & Fields, D. S. (1998) *Trends Biochem. Sci.* **23,** 109–113.
23. Robison, K., McGuire, A. M. & Church, G. M. (1988) *J. Mol. Biol.* **284,** 241–254.
24. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003) *Nucleic Acids Res.* **31,** 3497–3500.
25. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998) *Biological Sequence Analysis* (Cambridge Univ. Press, Cambridge, U.K.).
26. Egan, J. P. (1975), *Signal Detection Theory and ROC Analysis* (Academic, New York).
27. Ludwig, M. Z. & Kreitman, M. (1995) *Mol. Biol. Evol.* **12,** 1002–1011.
28. Wasserman, W. & Sandelin, A. (2004) *Nat. Rev. Genet.* **5,** 276–287.
29. Winfield, M. D. & Groisman, E. A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17162–17167.