Stochastic evolution of transcription factor binding sites

Johannes Berg and Michael Lässig*
Institut für Theoretische Physik, Universität zu Köln, Zülpicher Str. 77, 50937 Köln, Germany

A key step in the process of genetic transcription is the binding of one or several transcription factors to specific sites in the regulatory region of a gene. These binding sites may differ strongly across even closely related species, and the generation of new binding sites is an essential part of the evolution of regulatory networks. In this paper we consider the sequence evolution of binding sites, using empirically grounded fitness landscapes. We demonstrate how a new binding site for a given transcription factor may be generated *de novo* and estimate the time required for this process in terms of the neutral mutation rate, the selection coefficient, and the effective population size. We also consider how several sites binding to the same type of factor can co-exist in the regulatory region of a gene.

1. Introduction

The expression of a gene is regulated by products of other genes, termed transcription factors, as well as by external signals [1]. The molecular basis for this process is the physical binding of transcription factors to specific regions of DNA in the so-called regulatory region of a gene. Differences in gene regulation are believed to be a major source of diversity in higher eukaryotes. In this sense, suitable changes in the regulatory region of a gene may be viewed as a programming and reprogramming of the genetic network [2]. The driving force for such changes is evolutionary pressure. In this paper we consider the interplay between point mutations, selection, and genetic drift for transcription factor binding sites.

Binding sites in procaryotes consist of about $10-15$ base pairs relevant for binding and are found in most cases in the cis-regulatory region of a gene. In the model organism *E. coli*, the cis-regulatory region is about 300 base pairs long and contains a few transcription factor binding sites [3]. In any given regulatory region, there may be two or more sites binding the same factor.

The binding sequences for a given factor, both in the regulatory region of one gene as well as across different genes, are not identical to each other. The sequences constituting binding sites for the same transcription factor differ from each other by about $20-30$ of the relevant base pairs. This property is referred to as the *fuzziness* of binding sites, and makes the identification of binding sites from the regulatory regions of genes a challenging bioinformatical problem [4, 5, 6].

Regulation in eukaryotes is based on the same molecular mechanisms, but is vastly more complicated [7]. The cis-regulatory region is typically much longer (a few thousand base pairs) and contains a multitude of binding sites. At the same time, individual sites are shorter, with about 5-8 relevant base pairs. The sites are sometimes organized in *modules* interspersed between regions containing no sites.

The molecules acting as transcription factors may also physically interact with each other. Multiple binding sites involving different transcription factors and the interactions between them can be used to implement logical functions (such as transcription conditional on, say, the presence of two specific transcription factors and the absence of a third) [2]. Frequently one also finds multiple binding sites for the *same* transcription factor within the same regulatory region, suggesting that a single site may not be not be effective enough in binding the transcription factor.

The idea that binding sites and their sequences are at the heart of the molecular programming of regulatory networks puts the *evolution of binding sites* in the spotlight. The programming of genetic regulation

*Corresponding author. Email: lassig@thp.uni-koeln.de, Tel +49 221 470-3588, Fax +49 221 470-5159

must be achieved via the combined effects of mutation and selection, leading to new functions of genes as a response to specific demands. In fact one finds that over evolutionary time scales, binding sites can appear, disappear, or alter their sequence even between relatively closely related species; see, e.g., refs. [8, 9, 10, 11, 12]. This turnover of binding sites has been argued to follow an approximate molecular clock in *Drosophila* [13]. On the other hand, there are cases where binding sites are preserved even among fairly distant species. Both observations can be explained by evolution under selection pressure. Under constant external conditions, binding site sequences can be preserved over long periods (negative selection for change), while they respond quickly to new external conditions (positive selection for change).

In this paper, we focus mainly on the local sequence evolution of single binding sites. This is also the most promising starting point for a *quantitative* analysis of binding site evolution. We review a model of transcription factor binding, the *two-state model* by Berg and von Hippel [14], as well as the resulting fitness landscapes. We then discuss the stochastic modeling of the binding site evolution under point mutations, selection, and genetic drift. Details on these results can be found in [15]. In section 5 we consider the evolution of several binding sites for the same transcription factor in the regulatory region of a gene.

2. Factor binding and selection

The binding energy between a transcription factor and its binding site is, to a good approximation, the sum of independent contributions from a small number of important positions of the binding site sequence, $E = \sum_{i=1}^{\ell} \varepsilon_i$, with $\ell \approx 10 - 15$ [16]. The individual contributions ε_i depend on the position i and on the nucleotide a_i at that position. There is typically one particular nucleotide a_i^* preferred for binding; the sequence $(a_1^*, \ldots, a_\ell^*)$ is called the *target sequence*. Here we use the further approximation $\varepsilon_i = \varepsilon$ if $a_i = a_i^*$ and $\varepsilon = 0$ otherwise, the so-called *two-state model* [14]. The binding energy of any sequence $(a_1, \ldots, a_\ell)$ is then, up to an irrelevant constant, simply given by its Hamming distance r to the target sequence: $E = \varepsilon r$. (The Hamming distance is defined as the number of positions with a mismatch $a_i \neq a_i^*$.) The resulting binding probability of the factor in thermodynamic equilibrium is

$$p = \frac{1}{1 + \exp[\varepsilon(r - \rho)]}, \tag{1}$$

where ε is expressed in units of $k_B T$ and the offset term $\varepsilon\rho$ is a chemical potential. The parameters ε and ρ appropriate for typical binding sites have been discussed extensively in refs. [17, 18]. It is found that ε should take values around 2, which is consistent with measurements for known transcription factors giving $\varepsilon \approx 1 - 3$ [16, 19, 20]. The chemical potential depends on the number of transcription factors present in the cell, on the binding probability to *random* sites elsewhere in the genome (which have a sequence similar to the target sequence by chance), and on the binding to copies of the same operator other than the binding site in question. Binding to individual random sites is found to be negligible at the observed factor numbers n_f of about $50-5000$, which results in values $\rho \approx (\log n_f)/\varepsilon \approx 2 - 4$ [18]. Binding to other copies of the same operator becomes only relevant at low factor concentrations and high number of copies, when sites compete for factors.

These binding probabilities determine fitness landscapes for the binding site sequences. Following the conceptual framework of ref. [17], we assume that the environment of the gene to be expressed can be described by a number of *cellular states* (labelled by the index α) with different transcription factor concentrations, i.e., with different chemical potentials ρ^α. In each state, the fitness depends only on the expression level of the gene, which in turn is determined by the binding probability p^α of the transcription factor. Assuming that both dependencies are linear (this is not crucial) and that the states contribute additively to the overall fitness F, we obtain

$$F = \sum_\alpha s^\alpha p^\alpha, \tag{2}$$

where s^α is called the *selection coefficient* in the state α. Inserting (1), the fitness becomes a function of the Hamming distance r only.

In the simplest case, there are just two cellular states. The *on* state favours expression of the gene, the *off* state disfavours it. Assuming selection coefficients of equal magnitude $s = s^{\mathrm{on}} = -s^{\mathrm{off}}$, we obtain

a *crater* landscape,

$$F(r) = \frac{s}{1+\exp[\varepsilon(r-\rho^{\rm on})]} - \frac{s}{1+\exp[\varepsilon(r-\rho^{\rm off})]}, \tag{3}$$

with a high-fitness rim between $\rho^{\rm off}$ and $\rho^{\rm on}$ flanked by two sigmoid thresholds. If only the *on* state contributes significantly to selection, this reduces to the *mesa* landscape discussed in [17, 21],

$$F(r) = \frac{s}{1+\exp[\varepsilon(r-\rho^{\rm on})]}, \tag{4}$$

which has a high-fitness plateau of radius ρ and one sigmoid threshold. Hence, the parameters of the binding model have a simple geometric interpretation: ε gives the slope and the ρ^α give the positions of the sigmoid thresholds in the fitness landscape, see fig. 1(a,b). Clearly, the above is a minimal model of factor binding and its fitness landscapes, which neglects the context dependence of the binding process through cofactors, chromatin structure, and cooperative binding. However, it is a good starting point for order-of magnitude estimates of the adaptive evolution of binding sites.

3. Genetic drift and stochastic dynamics

The rates of nucleotide point mutations vary greatly between different organisms. In eukaryotes it is as low as $\mu \sim 2\times 10^{-9}$ in Drosophila [22]. The regime where a finite population evolves under stochastic fluctuations and selection is described by the Kimura-Ohta theory [23]. In this regime, mutants of fitness difference ΔF to an initially monomorphic population can substitute that population. This is a stochastic process, whose rate constant is given by

$$u = \mu N \, \frac{1-\exp(-2\Delta F)}{1-\exp(-2N\Delta F)} \tag{5}$$

in a diffusion approximation valid for $\Delta F \ll 1$ [24]. Here N is the *effective* population size (with an additional factor 2 for diploid populations). Eq. (5) has three well-known regimes. For substantially *deleterious* mutations ($N\Delta F \lesssim -1$), substitutions are exponentially suppressed. *Nearly neutral* substitutions ($N|\Delta F| \lesssim 1$) occur at a rate $u \approx \mu$ approximately equal to the rate of mutations in an individual. For substantially *beneficial* mutations ($N\Delta F \gtrsim 1$), the substitution rate is enhanced, with $u \simeq 2\mu N\Delta F$ for $N\Delta F \gg 1$.

In this picture, a population has a monomorphic majority for most of the time and occasional coexistence of two sequence states while a substitution is going on. The time of coexistence is $T \sim N$ for nearly neutral and $T \sim 1/\Delta F$ for strongly beneficial substitutions. The picture is thus self-consistent for $Tu \ll 1$, i.e., for $\mu N \ll 1$. Asymptotically, it describes monomorphic populations moving through sequence space with hopping rates u.

Introducing an *ensemble* of independent populations, this stochastic evolution takes the form of a Master equation. For a single binding site, we obtain

$$\begin{aligned}\tfrac{\partial}{\partial t}P(r,t) = \\ &(c-1)(\ell-r+1)\,u_{r-1,r}P(r-1,t) + \\ &(r+1)\,u_{r+1,r}P(r+1,t) - \\ &[r\,u_{r,r-1} + (c-1)(\ell-r)\,u_{r,r+1}]P(r,t).\end{aligned} \tag{6}$$

Here $P(r,t)$ denotes the probability of finding a population at Hamming distance r from the target sequence ($0 \le r \le \ell$, where ℓ is the length of the binding site), and $u_{r+1,r}$ is given by (5) with $\Delta F = F(r) - F(r+1)$. The combinatorial coefficients arise since a sequence at Hamming distance r can mutate in $(c-1)(\ell-r)$ different ways that increase r, and in r ways that decrease r, where $c = 4$ is the number of different nucleotides. The stationary distribution is

$$P_{\rm stat}(r) \sim \exp[S(r) + 2NF(r)]. \tag{7}$$

Here $S(r) = \log[\binom{\ell}{r}(c-1)^r/c^\ell]$ is the *mutational entropy* (the log fraction of sequence states with Hamming distance r) [21] and we have used the *exact result* $u_{r+1,r}/u_{r,r+1} = \exp\{2(N-1)\Delta F\}$. To derive (7) we then simply approximated $N-1$ by N. The form of $P_{\rm stat}(r)$ reflects the selection pressure, i.e., the scale s of fitness differences in the landscape $F(r)$. For near-neutral evolution ($2sN \ll 1$), $P_{\rm stat}(r) \sim \exp[S(r)]$ is simply a flat distribution on all sequence states. For moderate selection ($2sN \sim 1$), $P_{\rm stat}(r)$ results from a nontrivial balance of stochasticity and selection. For strong selection ($2sN \gg 1$), $P_{\rm stat}(r)$ takes appreciable values only at points of near-maximal fitness, where $F(r) \gtrsim F_{\rm max} - 1/2sN$. In this regime, the dynamics of a population consists of

beneficial mutations only, i.e., the system moves uphill on its fitness landscape.

4. Adaptive generation of a binding site

We now apply the dynamics (6) to the problem of adaptively generating a binding site in response to a newly arising selection pressure. We study a case of strong selection ($sN = 100$) in the crater fitness landscape (3) with parameters $\ell = 10$, $\varepsilon = 2$, $\rho^{\rm on} = 3$, $\rho^{\rm off} = 1$ (implying that the factor concentrations differ by a factor of 50), and a case of moderate selection ($sN = 7$) in the mesa landscape with parameters $\ell = 10$, $\varepsilon = 1$, $\rho = 3.6$. (The mesa type may be most appropriate for factors with multiple operator sites such as the CRP repressor in *E. coli*, where binding to an individual site is negligible in the *off* state.) The fitness landscapes for both cases are shown in fig. 1(a,b) in units of the selection pressure s. Substantially beneficial mutations occur only on their sigmoid slopes, i.e., in narrow ranges of r. The upper boundary of this region is given by $r_s = \rho^{\rm on} + \log[sN(\mathrm{e}^{\varepsilon} - 1)]/\varepsilon$, which takes typical values $r_s = 5{-}7$. In fig. 1(c,d), we show a sample history of adaptive substitutions from $r = 5$ to lower values of r, which are close to the point $r_{\max}$ of maximal fitness. The statistics of this adaptation is governed by the ensemble $P(r,t)$; the average $\overline{r}(t)$ and the standard deviation $\delta r(t)$ appear also in fig. 1(c,d). The expected time T_s of this adaptive process can be estimated by adding the expected times for each consecutive mutation towards a lower Hamming distance. In the case of strong selection, the expected time for such a mutation can be readily estimated in terms of the uphill rates in (6) and the expression for the fixation rate (5). Back-mutations towards a higher Hamming-distance are exponentially suppressed in this regime. One obtains

$$T_s = \frac{1}{2\mu N} \sum_{r=r_{\max}+1}^{r_s} \frac{1}{r(F(r-1) - F(r))}, \qquad (8)$$

taking values of a few times $1/s\mu N$.

Can such a selective sweep actually happen? This depends on the initial state of the regulatory region in question *before* the selection pressure for a new site sets in. The length of the regulatory region is denoted by L. The region is approximated as an ensemble of $L_1 = L - \ell + 1$ candidate sites undergoing *independent* neutral evolution, i.e., the simultaneous updating of ℓ sites by one mutation is replaced by independent mutations. At stationarity, the Hamming distance at a random site then follows the distribution $P_{\rm stat}(r) \sim \exp[S(r)]$ shown as empty bars in fig. 1(e,f). The minimal Hamming distance $r_{\min}$ in the entire region is given by the distribution $\mathcal{P}(r) = Q_{\rm stat}^{L_1}(r) - Q_{\rm stat}^{L_1}(r+1)$, where $Q_{\rm stat}(r) = \sum_{r' \geq r} P_{\rm stat}(r')$ is the cumulative distribution for a single site. $\mathcal{P}(r)$ is found to be strongly

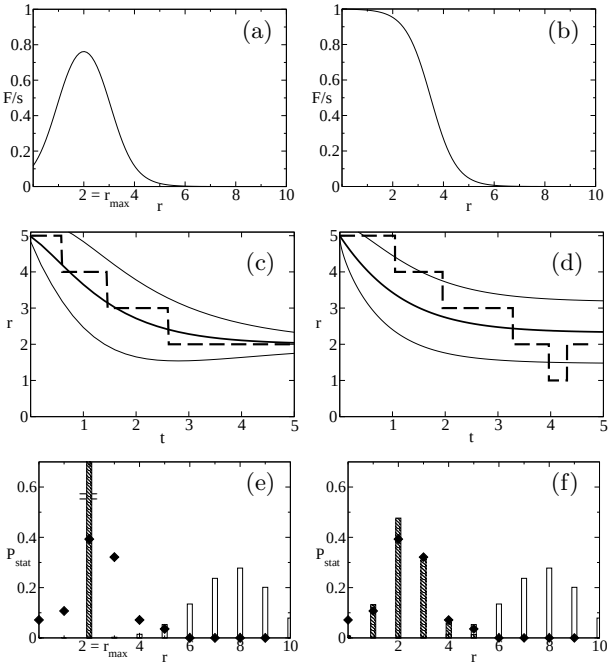

Fig. 1. Fitness landscapes and adaptive evolution for a single binding site. Strong selection ($sN = 100$, left column), moderate selection ($sN = 6.8$, right column). (a) *Crater* landscape (3). (b) *Mesa* landscape (4). (c,d) Adaptive dynamics as a function of time t measured in units of $1/2s\mu N$: Single history $r(t)$ (dashed lines), ensemble average $\bar{r}(t)$ (thick solid lines) and width given by the standard deviation curves $\bar{r}(t) \pm \delta r(t)$ (thin solid lines). (e,f) Stationary ensembles $P_{\rm stat}(r)$ of binding site sequences with selection (filled bars) and for neutral evolution (empty bars). Histogram of Hamming distances of CRP site sequences in *E. coli* from their consensus sequence (diamonds, from [17]).

peaked, taking appreciable values only in the range $\overline{r_{\min}}(\ell, L) \pm 1$ around its average. We assume the selective sweep sets in as soon as at least one site has a Hamming distance $r \leq r_s$. This is likely to happen spontaneously if $r_s \gtrsim \overline{r_{\min}}(\ell, L)$, leading to a joint condition on ℓ, L, and r_s. For $r_s \lesssim \overline{r_{\min}}(\ell, L) - 1$, there is a neutral waiting time before the onset of adaptation [15].

The stationary distribution $P_{\rm stat}(r)$ under selection is given by (7) and shown as filled bars in fig. 1(e,f). For strong selection, it is peaked at the point $r_{\max}$ of maximal fitness. For moderate selection, it takes appreciable values for $r = 0 - 4$: the binding site sequences are *fuzzy*. Assuming that the CRP sites at different positions in the genome of *E. coli* have to a certain extent evolved independently, we can fit $P_{\rm stat}(r)$ with their distance distribution (data taken from [17]). At the values of ε and $\rho^{\rm on}$ chosen, the two distributions fit well, see fig. 1(f).

Starting from a neutrally evolved initial state and progressing by point substitutions one can estimate the time for a selective sweep to generate a new site in response to a newly arising selection pressure. Such a selective sweep takes roughly $T_s \approx (\Delta r)/2s\mu N$ generations, where Δr is the number of adaptive substitutions required. For *Drosophila melanogaster*, with $\mu \approx 2 \times 10^{-9}$ [22] and $N \approx 10^6$, T_s is of the order of 10^7 generations or 10^6 years even for sites with a relatively small selection coefficient $s = 10^{-4}$. Such selective sweeps are faster than neutral evolution by a factor of about 100 and would allow for independent generation of sites even after the split from its closest relative *Drosophila simulans* about 2.5×10^6 years ago. Notice that new sites are more readily generated in large populations. As discussed above, generating a new site may also require a neutral waiting time until at least one candidate site in the regulatory region of the gene in question reaches the threshold distance r_s from the target sequence, where selection sets in. The expectation value of this neutral waiting time is termed T_0. For site formation to be efficient, however, selection must be able to set in spontaneously, i.e., T_0 must not greatly exceed the adaptive time T_s. This places a bound on the relevant length ℓ of the binding motif that can readily form in a regulatory region of length L. Given $L \approx 300$, for example, a motif with $\ell = 8$ and $r_s = 3$ could still allow for a spontaneous selective sweep. (For longer motifs, corresponding to groups of sites with fixed relative distance, this pathway would require regulatory regions of much larger L.) One may speculate that this adaptive dynamics is indeed one of the factors influencing the length of regulatory modules in higher eukaryotes.

For weaker selection, site fuzziness increases since $P_{\rm stat}$ extends beyond the sequence states of maximal fitness and is influenced by mutational entropy. As shown in fig. 1(f), one can explain in this way the observed fuzziness in CRP sites of *E. coli*. It would then reflect different evolutionary histories of independent populations, rather than sampling in one polymorphic population as in the quasispecies picture of refs. [17, 25]. However, the data are also compatible with strong selection if the selection coefficients s^α, and hence the value of $r_{\max}$, vary between different genes. Clearly, comparing $P_{\rm stat}$ with the distribution of sites in a single genome requires the assumption that the evolutionary histories of sites at different positions are at least to some extent independent. Future data of orthologous sites in a sufficient number of species will be more informative. Thus, further experimental evidence is needed to clarify the role of mutational entropy in the observed fuzziness.

5. Evolution of multiple binding sites

Regulation in higher organisms, where regulatory regions are several thousand base pairs long and often contain multiple binding sites, is characterized by the presence of *several* binding sites for a *single* type of transcription factor. The resulting fitness landscape depends in a complicated way on the Hamming distances of individual sites from the master sequence, as well as on collective properties such as the relative spacing of sites. Therefore, we will not attempt to construct a detailed model of selection here. We will ask a simpler question: assuming the function of a gene requires a given expression level J in the 'on' state, what are typical sequence configurations to be expected for multiple sites? We limit ourselves to the simplest case where (i) the total expression level generated by a group of sites is the sum of the binding probabilities at the individual sites and (ii) binding at a given sites does not depend on the occupation of other sites. The total expression level J for a group

of M sites with Hamming distances $r_1, \ldots r_M$ from the master sequence is then

$$J(r_1, \ldots, r_M) = \sum_{m=1}^{M} p(r_m) \ , \tag{9}$$

with $p(r)$ given by equation (1).

We now consider the equilibrium distribution $P_{\rm stat}(r_1, \ldots, r_M)$ under the constraint that the total current (9) remains constant. This can be obtained approximately by assuming that the contribution $p(r_m)$ of a given site to the total expression level depends on the time-averaged expression levels of all other sites,

$$P_{\rm stat}(r_1, \ldots, r_M) = \prod_{m}^{M} P_{\rm stat}(r_m) \tag{10}$$

with

$$P_{\rm stat}(r) \sim \exp\left\{S(r) + sp(r)\right\} \ . \tag{11}$$

Here the selection pressure s must be adjusted such that the required expression level J equals the expectation value given by (9) and (10),

$$J = \sum_{m}^{M} \langle p_m \rangle = M \langle p \rangle \tag{12}$$

with the average expression level contributed per site,

$$\langle p \rangle = \sum_{r=0}^{\ell} p(r) P_{\rm stat}(r) \ . \tag{13}$$

This mean-field approach neglects correlations between different sites. (Formally, it consists of a Legendre-transformation leading from a microcanonical to a canonical ensemble, a standard procedure in statistical mechanics.)

Figure 2(a) shows how the equilibrium distribution $P_{\rm stat}(r)$ changes with s for the binding probability (1) with $\rho = 1, \epsilon = 2$, and site length $\ell = 7$. At large values of s, all the weight of this distribution is concentrated at $r = 0$. As s decreases, a bimodal distribution emerges which a second peak near the maximum of the entropy $S(r)$. Thus only a certain fraction of the sites considered will actually bind the transcription factor. In the following, we consider a group of M potential binding sites. In this group we distinguish *active* sites, which actually bind the transcription factor, and *inactive* sites, which do not bind and in fact may have any sequence. A site is called an active site as long as $p(r) > 0.05\, p(r = 0)$, which gives a condition $r < r_0$ (In our particular example $r_0 = 3$. The particular choice of the threshold in the expression level has little influence on the results).

It is clear that the selection pressure s on an individual site decreases with increasing M. Indeed, the fuzziness observed experimentally suggests that the selection pressure on individual sites is rather low. Given the distribution $P_{\rm stat}$, the expected fuzziness of an active (i.e., observable) site is $\langle r \rangle_a \equiv \sum_{r \le r_0} r P_{\rm stat}(r) / \sum_{r \le r_0} P_{\rm stat}(r)$. We now use our model to predict the selection pressure s and the total number of sites, which is given by $M = J/\langle p \rangle$. Clearly, to maintain an expression level J, there must be at least $M_{\rm min} = J/p(r = 0)$ sites. This is the limit of no fuzziness and high selection pressure. As M increases, the selection pressure s decreases and the fuzziness $\langle r \rangle_a$ of active sites increases. Fig. 2(b) shows these quantities as a function of the ratio $x = M/M_{\rm min}$. At fixed J, the expected number of active sites remains approximately constant, $\langle M \rangle_a \approx M_{\rm min}(J) = M/x$.

As can be seen from Figure 2(b), already a moderate fuzziness (say $\langle r \rangle_a \approx 1/2$) corresponds to values of $x \approx 1/2$; i.e. the total number of sites M is about twice the number of active sites $\langle M \rangle_a$. A simple picture of the evolutionary dynamics emerges. At the observed levels of fuzziness, selection is too weak to ensure the conservation of the sites active at one point in time. Active sites will become defunct eventually due to deleterious mutations fixed by genetic drift. At the same time, other sites become active due to advantageous mutations. The stationary value of the expression level is maintained only at the level of the entire module.

6. Discussion

Transcription factors and their binding sites emerge as a suitable starting point for quantitative studies of gene regulation. Binding site sequences are short and their sequence space is simple. Moreover, explicit fitness landscapes can be derived from empirical data

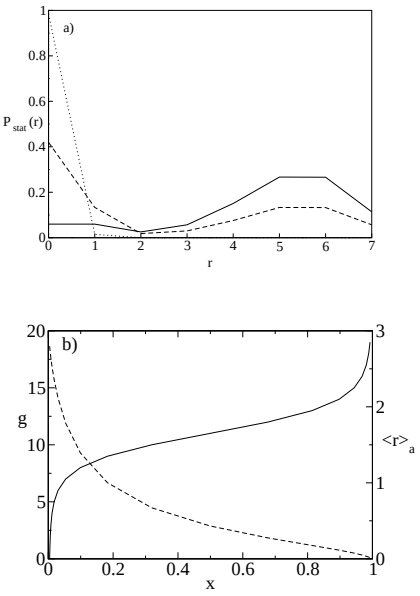

Fig. 2. (a) The equilibrium distribution $P_{\rm stat}(r)$ for $s = 4, 11, 20$ (solid, dashed, and dotted lines, respectively). (b) The selection pressure s and the average Hamming distance of active sites $\langle r\rangle_a$ as a function of $x = M_{\rm min}/M$ (solid and dashed lines respectively).

on binding affinities. For a single site, the simplest examples are of the *mesa* [17] or of the *crater* type, see fig. 1(a,b). For this case, the evolutionary dynamics of point mutations, selection, and genetic drift can be analyzed in some detail. The *de novo formation* of binding sites in response to an external change can be a rapid mode of evolution given even moderate selection pressures. Under neutral evolution, however, this mode would be too slow in many cases to account for the observed changes.

For the case of coexisting sites in eukaryotes, we have analyzed a simplified evolutionary model relating the collective properties of a module with multiple binding sites to the fuzziness and selection presssures of its constituent sites. Given typical levels of fuzziness found in observations, the model predicts low selection coefficient for each individual site and, hence, a considerable *turnover of sites*. That is, the number of sites observed in a single species is expected to be lower than the total number of sites observed over longer evolutionary times, e.g., by cross-species comparison.

In this picture, sites active at one point in time will tend to become inactive due to deleterious mutations, while other sites are (re-)activated due to compensatory selection. These large fluctuations in individual sites take place even if the regulatory module as a whole maintains a fairly constant expression level. In other words, compensatory selection can only be understood at the level of an entire module, not for its constituent sites. Hence, the evolution of a module is the *collective* dynamics of its sites. A consequence of our analysis is that bioinformatics methods identifying sites from inter-species sequence conservation will miss many functional sites. A more detailed analyis of multi-site modules and their evolutionary modes is a challenge for future research.

The present work was aimed at obtaining some insight into the molecular mechanisms and constraints underlying the dynamics of complex regulatory networks, thereby quantifying the notion of their *evolvability*. The programming of binding sites is found to provide efficient modes of adaptive evolution whose tempo can be quantified for the case of point mutations. The formation of complicated signal integration patterns and of multi-factor interactions, however, in higher eukaryotes requires generalizing our arguments in two ways. There are further modes of sequence evolution such as slippage events, insertions and deletions, large scale relocation of regulatory regions, and recombination. Moreover, there are also more general fitness landscapes describing, e.g., binding sites interacting via the expression level of the regulated gene (such as activator-repressor site pairs) and the coupled evolution of binding sites in different genes.

Acknowledgments

This work has been supported by DFG grant LA 1337/1-1.

References

[1] M. Ptashne and A. Gann. *Genes and Signals*. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY, 2002.

[2] N.E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA*, 100:5136–5141, 2003.

[3] J. Collado-Vides, B. Magasanik, and J.D. Gralla. Control site location and transcriptional regulation in Escherichia coli. *Microbiol. Reviews.*, 55:371–394, 1991.

[4] H.J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc. Nat. Acad. Sci. USA*, 97:10096–10100, 2000.

[5] G.Z. Hertz and G.D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

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

[7] J.R. Stone and G.A. Wray. Rapid evolution of cis-regulatory sequences via local point mutations. *Mol. Biol. Evol.*, 18(9):1764–1770, 2001.

[8] M.Z. Ludwig and M. Kreitman. Evolutionary dynamics of the enhancer region of even-skipped in drosophila. *Mol. Biol. Evol.*, 12(6):1002–1011, 1995.

[9] M.Z. Ludwig, N.H. Patel, and M. Kreitman. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125:949–958, 1998.

[10] E.T Dermitzakis, C.M Bergman, and A.G Clark. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.*, 20:703–714, 2002.

[11] J.L. Scemama, M. Hunter, J. McCallum, V. Prince, and E.Stellwag. Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. *J. Exp. Zool.*, 294:285–299, 2002.

[12] D. N. Arnosti. Analysis and function of transcriptional regulatory elements: Insights from Drosophila. *Ann. Review Entymology*, 48:579–602, 2003.

[13] J. Costas, F. Casares, and J. Vieira. Turnover of binding sites for transcription factors involved in early Drosophila development. *Gene*, 310:215–220, 2003.

[14] O.G. Berg and P.H. von Hippel. Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol.*, 193:723–750, 1987.

[15] J. Berg, M. Lässig, and S. Radic. On the evolution of transcriptional regulation. submitted.

[16] D.S. Fields, Y. He, A.Y. Al-Uzri, and G.D. Stormo. Quantitative specificity of the mnt repression. *J. Mol. Biol.*, 271:178–194, 1997.

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

[18] U. Gerland, D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc.Nat. Acad. Sci. USA*, 99:12015–12020, 2002.

[19] M. Oda, K. Furukawa, K.Ogata, A. Sarai, and H. Nakamura. Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain. *J. Mol. Biol.*, 276:571–590, 1998.

[20] A. Sarai and Y. Takeda. RT lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc. Nat. Acad. Sci. USA*, 86:6513–6517, 1989.

[21] L. Peliti. Quasispecies evolution in general mean-field landscapes. *Europhys. Lett.*, 57:745–751, 2002.

[22] C. Schlötterer, M.-T. Hauser, A. v. Haeseler, and D. Tautz. Comparative evolutionary analysis of rDNA ITS regions in drosophila. *Mol. Biol. Evol.*, 11:513–522, 1994.

[23] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61:763–771, 1969.

[24] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

[25] A. Sengupta, M. Djordjevic, and B. Shraiman. Specificity and robustness in transcription control networks. *Proc. Nat. Acad. Sci. USA*, 99:2072–2077, 2002.