Supporting Material to: Physical limits on cooperative protein-DNA binding and the kinetics of combinatorial transcription regulation

Nico Geisel
Departament de Fisica Fonamental, Facultat de Fisica,
Universitat de Barcelona, Barcelona, Spain

Ulrich Gerland
Arnold-Sommerfeld Center for Theoretical Physics
and Center for Nanoscience (CeNS),
Ludwig-Maximilians-Universität, München, Germany

September 17, 2011

S1 Exact calculation of steady-state activities

Single TF molecules. We first treat the case where the cell contains only a single molecule of each TF species, $N_A = N_B = 1$. The equilibrium statistics of the system is described by the canonical ensemble of statistical physics. The appropriate Boltzmann weight for a single TF binding to one of $L_{\rm G}$ sites in a non-specific DNA background is $q_{ns} = \exp(-E_{\rm ns})$ (see below for the most general case with an arbitrary background and larger TF numbers). For a purely non-specific background and $S = V_{\rm cell}/V_{\rm TF} \gg L_{\rm G}$ unbound states, the partition function is

$$\begin{aligned} Z_{back} &= L_{\rm G}(L_{\rm G} - 2L)q_{\rm ns}^2 + S^2 \\ &+ 2SL_{\rm G}q_{\rm ns} \\ &+ \omega\left(L_{\rm G}q_{\rm ns}^2 + S\right) . \end{aligned} \tag{S1}$$

The first three terms describe the non-interacting states, where A and B are either separately bound to the DNA to non-adjacent sites, or both are free but not dimerized, or one is DNA-bound and the other is free. The fourth term corresponds to the states where A and B are dimerized, either on the DNA or unbound. The fraction of dimers in the background corresponds to the ratio of the weights of the dimerized states to the weight of all possible states, $\omega\left(L_{\rm G}q_{\rm ns}^2 + S\right)/Z_{back}$. Rewriting this expression in terms of the monomer DNA binding ratio $\alpha = P_d/P_c = q_{\rm ns}L_{\rm G}/S$, one obtains

$$P_{\rm dimer}(\alpha,\omega) = \frac{\omega}{\omega + \left(S(\alpha^2+1) + 2\alpha\right)/(\alpha q_{\rm ns} + 1)} . \tag{S2}$$

For a binding ratio of one, i.e. when the monomers are optimized for independent search, $P_{\rm dimer}(\omega) = \omega/(\omega+2L_{\rm G})$, which is the case plotted in Fig. 3A. Here, a dimerization probability of 0.5 is reached at $\omega_{1/2} = 2L_{\rm G}$, while we would have $\omega_{1/2} = S$ for $\alpha \to 0$ and $\omega_{1/2} = L_{\rm G}$ for $\alpha \to \infty$.

Eq. S1 provides the binding-statistics on non-target states. To study the full system, we add the target states with weights $q_T = \exp(-E_T)$ for the full partition function

$$Z_{tot} = Z_{back} + [2(L_{\rm G} - L - 1)q_{\rm ns} + S]q_T + \omega q_T^2 , \tag{S3}$$

where the second term is the weight of a single occupied target and the third term is the weight for both targets to be occupied simultaneously. Hence the double target occupation probability is $p_{ab} = \omega q_T^2/Z_{tot}$. This equation can be solved for q_T at given values of p_{ab} and ω (since our analysis assumes a fixed p_{ab} corresponding to the optimal occupation-probability of the targets in the ON-state). Hence we obtain an explicit expression for $E_T\left(\omega, p_{ab}\right)$ (not shown), which we use throughout this paper to determine E_T for the kinetic model and stochastic simulations in the $N = 1$ case. Furthermore, to calculate the fold-change $\phi = p_{ab}/p_a$ at a given $E_T\left(\omega, p_{ab}\right)$ we determine the probability of single TF target binding p_a in the absence of a partner. By calculating the partition function for a system of a single TF, we find

$$p_a = p_b = \frac{{\rm e}^{-E_T\left(\omega,\, p_{ab}\right)}}{(L_{\rm G} - 1)q_{\rm ns} + S + {\rm e}^{-E_T\left(\omega,\, p_{ab}\right)}} . \tag{S4}$$

For small ω, this probability scales as $\sim \omega^{-1/2}$.

Multiple TF molecules. For the case of multiple TF molecules, we calculate the exact equilibrium statistics of our full model using the standard transfer matrix approach from statistical physics, see e.g. (1, 2). The calculation is based on the grand canonical ensemble, i.e. the average copy numbers N_A, N_B of the proteins A and B are set by the corresponding chemical potentials μ_A , μ_B . The total partition function Z of the complete system then factorizes,

$$Z = Z_d Z_c \ , \tag{S5}$$

into a product of a "DNA partition function" Z_d involving only the DNA-bound states of the TFs and a "cytosol partition function" Z_c involving only the unbound states (the factorization is possible because DNA-bound TFs do not interact with unbound TFs and because the TF numbers are not conserved in the grand canonical ensemble). Due to the low TF concentrations in the cytosol, steric exclusion between unbound TFs is negligible, and Z_c takes the simple form

$$Z_c \ = \ \left(1 + \mathrm{e}^{\mu_A} + \mathrm{e}^{\mu_B} + \omega\, \mathrm{e}^{\mu_A + \mu_B}\right)^S \ , \tag{S6}$$

where $S \gg L_\mathrm{G}$ is the number of solvent states (i.e. the ratio of the cell volume to a characteristic TF volume, $S = V_\mathrm{cell}/V_\mathrm{TF}$) and the statistical weight for an unoccupied solvent state is one. For the calculation of the DNA partition function Z_d, we do take the steric exclusion of DNA-bound TFs into account. The number of base pairs covered by a single TF molecule is denoted by L. Each base pair $i = 1 \ldots L_\mathrm{G}$ on the genome can then be in one of $2L + 1$ states: In state 0, the base pair is not covered by a TF. In state 1, it is the leftmost contact position of a TF of type A, in state 2 it is the second leftmost contact position, and so on, up to state L corresponding to the rightmost contact position of A. States $L + 1$ up to $2L$ are analogous for B. The transfer matrix Q_i describes the statistical coupling between the states of the neighboring DNA positions i and $i + 1$. Each Q_i is a square matrix of dimension $2L + 1$, defined such that the partition function is equal to the trace of the (ordered) product of all transfer matrices,

$$Z_d = Tr\left(\prod_{i=1}^{L_\mathrm{G}} Q_i\right) \ , \tag{S7}$$

for a circular DNA with L_G basepairs (for a linear DNA molecule, the trace operation would have to be replaced by multiplication of a row vector from the left and a column vector from the right, with the vector components properly chosen to enforce the boundary conditions). Let us denote by $[Q_i]_{ss'}$ the element in row s and column s' of the transfer matrix at position i. It takes on a non-negative value, which corresponds to the conditional statistical weight of finding position i in state s', provided that position $i - 1$ is in state s. Thus, each $[Q_i]_{ss'}$ is a Boltzmann factor that accounts for the contribution to the total configurational energy that stems from position i and its interaction with position $i + 1$. The Boltzmann factor is zero, if the two states are incompatible (overlapping TFs or a single TF binding to non-contiguous basepairs). The non-zero entries of Q_i contain the protein-DNA binding energy landscapes E_i^A and E_i^B , the cooperativity ω, and the

chemical potentials. For illustration, we show the transfer matrix Q_i for TFs of length $L = 2$,

$$Q_i = \begin{pmatrix} 1 & \mathrm{e}^{-E_i^A+\mu_A} & 0 & \mathrm{e}^{-E_i^B+\mu_B} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & \mathrm{e}^{-E_i^A+\mu_A} & 0 & \omega\,\mathrm{e}^{-E_i^B+\mu_B} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & \mathrm{e}^{-E_i^A+\mu_A} & 0 & \mathrm{e}^{-E_i^B+\mu_B} & 0 \end{pmatrix} . \tag{S8}$$

The entries with value one reflect the mere compatibility of neighboring states without an energetic contribution (e.g., when position $i-1$ is in state 1, position i must be in state 2, and there is no additional energy contribution to take into account). Note that we assume a directional interaction between the TFs A and B (the attractive contact only occurs when B is bound directly downstream from A).

From the partition function (S5), we can obtain exact expressions for the occupation probabilities of DNA sites by differentiation. For instance, the probability that a TF molecule of type A is bound to the site starting at position i on the DNA is

$$p_i^A = -\frac{\partial}{\partial E_i^A} \log Z \ . \tag{S9}$$

The derivative is straightforward to evaluate explicitly, leading to an expression of the form $p_i^A = Z'_d/Z_d$, where the restricted partition function Z'_d has the same form as (S7), but with a projection matrix next to Q_i inside the trace. This exact expression is easily computed numerically, in particular when large parts of the binding energy landscapes E_i^A and E_i^B are flat (equal to the non-specific binding energy E_{ns}), since large parts of the product in (S7) then reduce to matrix powers (which are quickly calculated via diagonalization). Similarly, the probability of cooperative binding at site i is calculated starting from the expression

$$p_i^{AB} = \frac{\partial^2}{\partial E_i^A\, \partial E_{i+L}^B} \log Z \ , \tag{S10}$$

where the derivatives enforce that a B molecule is bound directly adjacent to the A molecule, such that together they cover the DNA positions from i to $i+2L-1$. Finally, the average number of TF molecules in the system at given values of the chemical potentials μ_A, μ_B are obtained by summing over the occupation numbers of all states, e.g.

$$N_A = \sum_{i=1}^{L_{\mathrm{G}}} p_i^A + \frac{S(\mathrm{e}^{\mu_A} + \omega\,\mathrm{e}^{\mu_A+\mu_B})}{Z_c} \ . \tag{S11}$$

Similarly, the average number of dimers in the system is

$$N_{\mathrm{dimer}} = \sum_{i=1}^{L_{\mathrm{G}}} p_i^{AB} + \frac{S\,\omega\,\mathrm{e}^{\mu_A+\mu_B}}{Z_c} \ , \tag{S12}$$

from which the fraction of dimers, $P_{\mathrm{dimer}}(\omega) = N_{\mathrm{dimer}}/N$, in Fig. S2A is computed. The fold-

change ϕ in Fig. S2B is calculated as the ratio of the dimer occupancy (S10) at the target site pair in the presence of both TFs ($\mu_A = \mu_B \equiv \mu$ such that $N_A = N_B \equiv N$) to the monomer occupancy (S9) at its target site when only one TF is present (μ_A chosen such that $N_A = N$ while μ_B is set to a large negative value such that $N_B \approx 0$).

The above framework can be used to calculate any equilibrium observable exactly for our full model and it also provides a reference point for our kinetic simulations, which produce equilibrium values in the long-time average. However, it is also useful to derive a simple approximation to the exact solution of the multiple TF molecule case, which still incorporates the effect of a (nonspecific) DNA background, but neglects steric exclusion between the TFs in the background. Assuming $\mathrm{e}^{E_{\mathrm{ns}}} \ll 1$ and $N \ll L_{\mathrm{G}}$, and again taking a DNA binding ratio of one, such that $S\mathrm{e}^{E_{\mathrm{ns}}} = L_{\mathrm{G}}$, we find

$$q_{\mathrm{ns}} \equiv \mathrm{e}^{-E_{\mathrm{ns}}+\mu} \approx \frac{\sqrt{1+\frac{N}{L_{\mathrm{G}}}\omega}-1}{\omega}\,, \tag{S13}$$

which leads to the background dimerization fraction

$$P_{\mathrm{dimer}}(\omega) \approx 1 - \frac{2L_{\mathrm{G}}}{N\omega}\left(\sqrt{1+\frac{N\omega}{L_{\mathrm{G}}}}-1\right) \tag{S14}$$

that we use in Eq. 4 of the main text for the approximative form of the cooperative search time.

S2 Stochastic simulation of cooperative search kinetics

To study the cooperative search process within the full reaction scheme of Fig. 2B, we implemented a kinetic Monte Carlo simulation based on the standard Gillespie algorithm. For our simulations, we used fixed numbers, N_A and N_B, of A and B molecules (i.e., any equilibrium values computed in these simulations correspond to thermodynamic averages in the canonical ensemble). The state of the system is specified by the state of each TF molecule, which can be either free or dimerized in solution, or bound to the DNA at position p. The simulations generate stochastic continuous-time trajectories in this discrete state space. Each simulation step consists of one of the moves depicted in Fig. 2B, however the set of available moves depends on the current state of the system. In particular, moves that would violate the steric constraint that each DNA basepair can be be in contact with only a single TF molecule cannot be chosen. Thus, TF molecules can, for instance, not change the order at which they are bound along the DNA solely via sliding moves.

To measure the average cooperative search time $\langle\tau\rangle$, we perform 100 simulations for each set of model parameters. Each simulation run is initialized in the state where all molecules are unbound (this mimics the condition of a cell prior to receiving a signal that triggers allosteric activation of TF-DNA binding), and terminated once the the two adjacent target sites are both occupied simultaneously. The data points in Fig. 3C, Fig. 4, and Fig. S2C correspond to the simulation time averaged over the 100 runs. Another observable of interest here is the relative contribution of the dimer pathway to the search process, as shown in Fig. 3D and Fig. S2D. This observable corresponds to the fraction of simulation runs where the final state is reached by a dimer move, such that both targets simultaneously become occupied by their cognate TF molecule.

S3 Analytical description of the cooperative search kinetics

Here, we develop a simplified analytical description of the cooperative search kinetics, which distinguishes only the target occupation states and the two search modes (dimeric vs. monomeric). As shown in Fig. S1, this description corresponds to a kinetic scheme with four states and six effective rates. The scheme amounts to two competing Michaelis-Menten type processes which lead to the same final state. The initial state 2 corresponds to the state of our TF-DNA system where both proteins are unbound. From there, the target state can either be reached via state 1 (dimer pathway) or via state 3 (monomer pathway). The dimer pathway is kinetically characterized by the effective dimerization rate r_2^-, the effective dissociation rate r_1^+, and the dimer search rate $r_1^- \equiv 1/\langle \tau_D \rangle$. Similarly, the monomer pathway is characterized by the three rates r_2^+, r_3^-, and r_3^+. Since state 3 does not distinguish whether A or B is bound, the rate $r_2^+ \equiv 2/\langle \tau_M \rangle$ is twice the monomer search rate. In contrast, the rate $r_3^+ \equiv 1/2\langle \tau_M \rangle$ corresponds to only half the search rate of a monomer because one target is already occupied and the other target is accessible from one side only. Finally, r_3^- is the total rate at which a monomer dissociates from its target, either via sliding or unbinding.

We can express the three remaining undetermined rate constants r_2^-, r_1^+, and r_3^- in terms of our underlying model parameters. For arbitrary binding energy landscapes, the effective dimerization rate is

$$r_2^- = \sum_{i \neq a,b} \left[\left(k_i^{A+} + k_{i+L+1}^{B-} \right) p_i^A \, p_{i+L+1}^B + k_a \, p_i^A P_c^B + k_a \, p_i^B P_c^A \right] + k_a P_c^A P_c^B \ , \tag{S15}$$

where we have used the equilibrium probabilities introduced above in section A of 'Methods', and P_c^A, P_c^B denote the equilibrium probabilities for the TFs to be unbound in solution. The rates k_i^{A+} and k_i^{A-} denote the forward and backward sliding rates from position i, see section 'Full model'. Using our approximations from section A for a non-specific background, we find the simpler form for the effective dimerization rate

$$r_2^- = \left(\frac{2\, k_{\rm sl}}{L_{\rm G}} - k_a \right) P_d^2 + k_a \ , \tag{S16}$$

where $P_d = 1 - P_c^A = 1 - P_c^B$ is the probability to find a TF molecule bound to DNA. Similarly, the effective dissociation rate has the general form

$$r_1^+ = \sum_{i \neq a} \frac{p_i^{AB}}{\omega} \left(k_i^{A,\rm off} + k_{i+L}^{B,\rm off} + k_i^{A-} + k_{i+L}^{B+} \right) + k_{\rm d} \, P_c^{AB} \ , \tag{S17}$$

where $k_i^{A,\rm off}$ denotes the site-specific DNA-unbinding rate for A and P_c^{AB} is the probability to find the two TFs dimerized in solution. The simplified effective dissociation rate for a non-specific background is

$$r_1^+ = \frac{2 P_d^{AB}}{\omega} \left(k_{\rm sl} + k_{\rm off} \right) + k_{\rm d} \, P_c^{AB} \ , \tag{S18}$$

where P_d^{AB} is the total probability to find the TFs non-specifically bound to the DNA as a het-

erodimer. Finally, the total rate for monomer loss from a target is

$$r_3^- = k_{\mathrm{off},a} + 2\,k_{\mathrm{sl},a}\ . \tag{S19}$$

where the index a indicates that these are unbinding and sliding rates from the target site, which are slower than their bulk counterparts by the additional Boltzmann factor corresponding to the energy difference between the non-specific binding energy and the target binding energy, see section 'Full model'.

With these rates, the average assembly time of the two TFs on the double target corresponds to the mean first passage time (MFPT) of a random walker hopping between the four sites at the given site-dependent jump rates. The random walker starts at site 2 and terminates on the target site. We use the standard MFPT formalism as described, for instance, in Ref. (3) to calculate this cooperative search time. The general formula for the MFPT $\langle\tau(M)\rangle$ starting from site M on a linear lattice with $N+1$ sites, with the two boundary sites 0 and N both absorbing, is

$$\langle\tau(M)\rangle = W(M)\sum_{m=1}^{N-1}\sum_{n=1}^{m}\frac{1}{r_n^+}\prod_{j=n+1}^{m}\frac{r_j^-}{r_j^+} - \sum_{m=1}^{M-1}\sum_{n=1}^{m}\frac{1}{r_n^+}\prod_{j=n+1}^{m}\frac{r_j^-}{r_j^+}\ , \tag{S20}$$

where $W(M)$ is the total probability to exit to site N,

$$W(M) = \frac{1+\sum_{m=1}^{M-1}\prod_{j=1}^{m}\frac{r_j^-}{r_j^+}}{1+\sum_{m=1}^{N-1}\prod_{j=1}^{m}\frac{r_j^-}{r_j^+}}\ . \tag{S21}$$

For the problem at hand, we have $N=4$ and $M=2$. Defining the Michaelis-Menten-type constant $K_1 = (r_1^- + r_1^+)/r_2^-$ for state 1 and $K_3 = (r_3^+ + r_3^-)/r_2^+$ for state 3, we can rewrite the cooperative search rate, i.e. the inverse average search time, in the compact form

$$\frac{1}{\langle\tau\rangle} = \frac{K_1\,r_3^+ + K_3\,r_1^-}{K_1 + K_1K_3 + K_3}\ , \tag{S22}$$

which is the expression used to obtain the lines in Fig. 3C. In the limit where r_2^- vanishes, this reduces to the average search rate for two independent monomers,

$$\frac{1}{\langle\tau_{A,B}\rangle} = \frac{r_3^+}{1+K_3}\ . \tag{S23}$$

Using the relation $2\,r_3^-\,p_a = r_2^+(1-p_a)$, we can rewrite the corresponding search time in the form

$$\langle\tau_{A,B}\rangle = \left(\frac{5}{2} + \frac{1-p_a}{p_a}\right)\,\langle\tau_M\rangle\ , \tag{S24}$$

which best explains the effect of missed encounters where $1/p_a$ is the average number of times a TF must return to the target before finding the other target occupied. In the small ω regime the cooperative search process corresponds to an independent monomer search and $\langle\tau\rangle \approx \langle\tau_{A,B}\rangle$.

Given that $p_a \sim \omega^{-1/2}$, this form also explains the $\langle\tau\rangle \sim \sqrt{\omega}$ scaling of the search time at small cooperativities.

We can further simplify Eq. S22 by noting that the average search time is virtually identical (in the parameter regime considered here) when the search begins in state 1 instead of state 2. With state 1 as the initial state, we find

$$\langle\tau\rangle = \left(r_1^- P_{\rm dimer} + \frac{1}{\langle\tau_{A,B}\rangle}(1 - P_{\rm dimer}) \right)^{-1} . \tag{S25}$$

The first term corresponds to the dimer pathway, while the second term corresponds to the monomer pathway. As expected, the contribution of either pathway depends on the dimerization probability and on the search rate of the respective mode. It follows that the relative weight of the dimer pathway can be written as

$$W_D(\omega) = \frac{P_{\rm dimer}(\omega)\, r_1^-}{P_{\rm dimer}(\omega)\, r_1^- + (1 - P_{\rm dimer}(\omega))\langle\tau_{A,B}\rangle^{-1}} , \tag{S26}$$

which was used to obtain the lines in Fig. 3D. It is straightforward to generalize these equations also to the case of $N > 1$, where the dimerization probability $P_{\rm dimer}(\omega, N)$ becomes a function of both ω and N, and the search rate for each mode increases by a factor of N: $r_1^- \to N r_1^-$ and $\langle\tau_{A,B}\rangle \to \langle\tau_{A,B}\rangle/N$. In this case we obtain Eq. 4 from the main text which is used to obtain the analytical curves in Fig. S2C. Using the dimerization probability $P_{\rm dimer}(\omega, N)$, we also extend Eq. S26 to the case of $N > 1$, to obtain the curves in Fig. S2D.

S4 Additional notes

To obtain an estimate of the number of *E. coli* operons which are regulated by two or more transcription factors, we perused the "RegulonDB" database (4). At the time of writing, this database lists 370 *E. coli* operons as regulated by a single transcription factor, while 383 operons are listed as regulated by two or more transcription factors (188 of these are believed to be regulated by exactly two transcription factors).

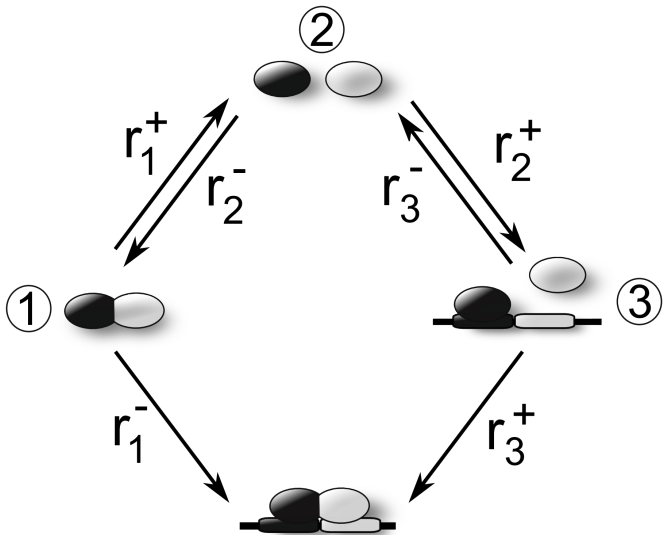

Figure S1: Simplified model used to calculate the mean cooperative search time analytically. In this model only the different target occupation states and the dimeric vs. monomeric search modi are distinguished. The rates r_1^- and r_3^+ correspond to the search rates of dimers or monomers respectively, whereas r_2^- and r_1^+ are the total rates at which a dimerization or a dissociation occur in the dimeric or monomeric state, respectively. The rate r_3^- refers to the total rate at which a monomer leaves its target, either by sliding away or by dissociating from it.

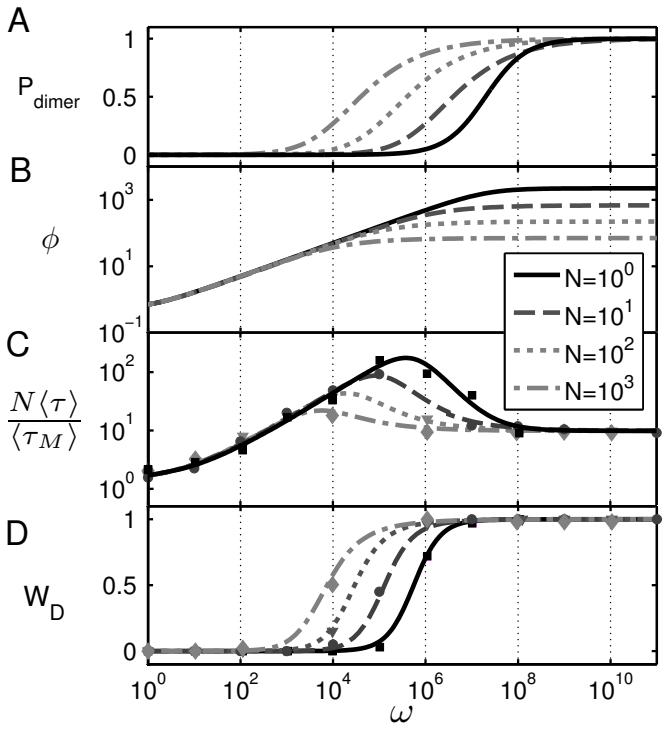

Figure S2: Cooperative search times and steady state levels as a function of ω, given different TF copy numers $N = 1, 10, 100$, and 1000 at a fixed $p_{ab} = 0.5$. (A) The dimerization threshold decreases with increasing TF concentrations whereas the foldchange (B) is independent of the TF number in the monomeric regime. The maximal foldchange is reached at the dimerization threshold, which decreases with the TF concentration, such that the maximal foldchange in (B) decreases as well. The search time (C) scales as $1/N$ in the purely monomeric and purely dimeric regime. In the intermediate regime, the maximal search time decreases stronger than $1/N$, as the onset of the dimeric pathway (shown in D) moves to lower cooperativities.

References

1. Schwabl, F., 2006. Statistical Mechanics. Springer, 2nd edition.

2. Teif, V. B., 2007. General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to OR operator in phage lambda. *Nucleic Acids Res.* 35:e80.

3. Gardiner, C., 2004. Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. Springer, 3rd edition.

4. Gama-Castro, S., H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides, 2011. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* 39:D98–D105.