Evolution of DNA motifs by protein binding

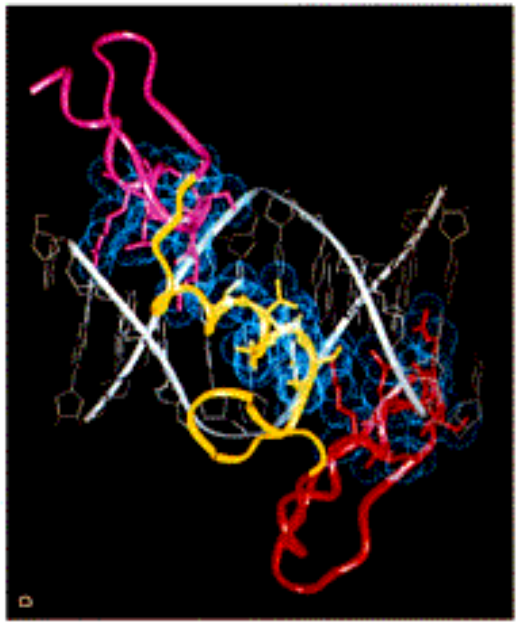
U. Gerland & T. Hwa

Department of Physics, University of California at San Diego

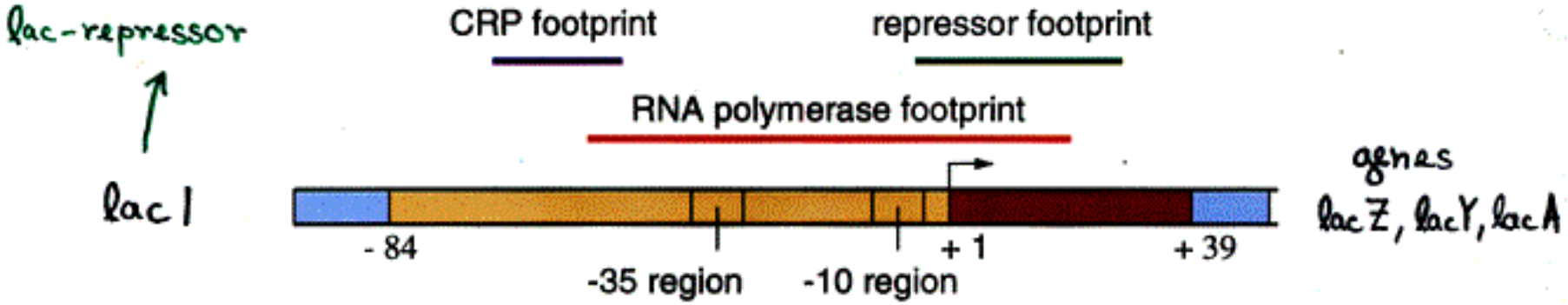
Outline

- Review of Protein-DNA interaction
- Evolutionary model for a single binding site
- Selection threshold and stationary distributions
- Evolution of a pair of sites
- Application to CRP binding sites in *E. coli*

Acknowledgments: Hochschulsonderprogramm III of the DAAD (U.G.),
NSF Grant No. DMR-9971456, DBI-9970199, Beckmann Foundation



- consider only passive protein-DNA binding (no ATP consumption)
- transcription factors bind to specific operator sites on DNA
- hydrogen bonds + electrostatic interaction
- transcription factors are
  activators or repressors
- Example: regulation of *lac* - Operon



AAAGTGTGACGCCGTGCAAATAAT
ATTTCGTGATGTTGCTTGCAAAAA
TTTATGTGCGCATCTCCACATTAC
ATTCTGTAACAGAGATCACACAAA
CCTTTGTGATCGCTTTCACGGAGC
AAAACGTGATCAACCCCTCAATTT
AACTTGTGGATAAAATCACGGTCT
GTTTTGTTACCTGCCTCTAACTTT
TTAATTTGAAAATTGGAATATCCA
AATTTGCGATGCGTCGCGCATTTT
TTAATGAGATTCAGATCACATATA
AATGTGTGCGGCAATTCACATTTA
GAAACGTGATTTCATGCGTCATTT
AAATGACGCATGAAATCACGTTTC
TTGCTGTGACTCGATTCACGAAGT
TTTTTGTGGCCTGCTTCAAACTTT
GAATTGTGACACAGTGCAAATTCA
ATAATGTTATACATATCACTCTAA
CGATTGTGATTCGATTCACATTTA
GTTTTGTGATGGCTATTAGAAATT
GAACTGTGAAACGAAACATATTTT
AATGTGTGTAAACGTGAACGCAAT
TTTGTGTGATCTCTGTTACAGAAT
GTAATGTGGAGATGCGCACATAAA
TTTTTGCAAGCAACATCACGAAAT
TTAATGTGAGTTAGCTCACTCATT
ATTATTTGCACGGCGTCACACTTT
ATTATTTGAACCAGATCGCATTAC
TAATTGTGATGTGTATCGAAGTGT

CRP binding sites in *E.Coli*

Observations:

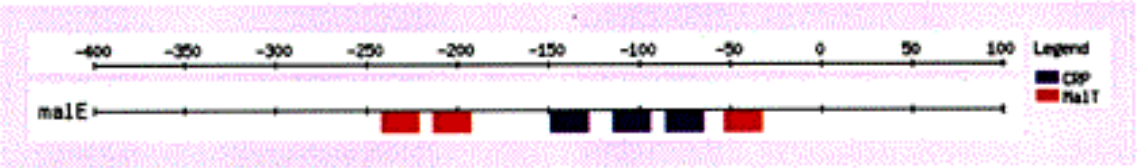
- 10 nucleotides carry most of the specificity
- large variation around consensus TGTGA...TCACA

⇒ fuzzy motifs, i.e. no DNA-protein recognition code

- sites often occur in doublets or triplets

Hypothesis:

Sequences result from an equilibrium between mutational drift and selection pressure towards high binding affinity



Review: Protein-DNA binding

Binding energy:

independent contribution from each nucleotide

$$E_{\vec{S}} \approx \sum_{i=1}^{L} E_i(S_i)$$

Simplification:

2-state model

$$E_i(S_i) = \begin{cases} 0\,, & \text{match} \\ \varepsilon\,, & \text{mismatch} \end{cases}$$

take only most significant positions into account

Typical values:

$L = 10 \ldots 15$, $\varepsilon = 1 \ldots 3\, k_B T$

Binding energies for Mnt (in $k_B T$):

pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(D.S. Fields, Y. He, A.Y. Al-Uzri & G. Stormo, 1997)

Binding probability:

depends only on number $r_{\vec{S}} = |\vec{S} - \vec{S}^*|$ of mismatches from optimal binding sequence $\vec{S}^*$

$$P(r) = \frac{1}{1 + e^{\varepsilon(r - r_0)}} \qquad 1 \le r_0 \le 5$$

$\varepsilon\, r_0$ = chemical potential (depends on [protein] and non-specific binding)

Goals

- Quantify the selection pressure needed for the maintenance of motifs
- Derive nucleotide statistics at given selection pressure and mutation rate
  → explain fuzziness
- Estimate selection pressure acting on an individual binding site
- Explain and interpret the occurrence of doublets and triplets

Evolutionary Model

Binding sequence: $\vec{S} = \{S_1, S_2, ..., S_L\}$ $\longleftarrow$ Selection + Mutation

L nucleotides $S_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$

Selection:

"fitness" = reproduction rate $\Phi_{\vec{S}}$

no binding: $\Phi_{\vec{S}} = \phi_0 = \ln 2/\tau$

perfect binding: $\Phi_{\vec{S}} = \phi_0(1+\alpha)$

$\Rightarrow \Phi_{\vec{S}} = \phi_0 \cdot (1 + \alpha \cdot P_{\vec{S}})$

with $P_{\vec{S}}$ = binding probability

Mutation:

single-nucleotide mutations

$S_i \to S_j \neq S_i$ at rate μ_0

total mutation rate of sequence

$\mu = \mu_0 \, (\mathcal{A} - 1) \, L$

typical ($E.coli$):

$\mu \sim 10^{-8}/\tau, \quad \tau \approx 10\,\text{min}$

α = dimensionless selection pressure: $\begin{cases} \alpha \to 0 & \Leftrightarrow \text{ no selection} \\ \alpha \to \infty & \Leftrightarrow \text{ no binding is lethal} \end{cases}$

$N_{\vec{S}}(t) = \langle$ # of individuals with sequence $\vec{S}\,\rangle \to$ Mean-field equation

$$\partial_t N_{\vec{S}}(t) = \mu_0 \sum_{\vec{S}'} N_{\vec{S}'}(t)\, \delta_{|\vec{S}-\vec{S}'|,1} - \mu\, N_{\vec{S}}(t) + \Phi_{\vec{S}}\, N_{\vec{S}}(t)$$

"Radial" evolution equation

Fitness depends only on $r \Rightarrow$ introduce radial distribution

$$N(r,t) = \sum_{\vec{s}} N_{\vec{s}}(t)\, \delta_{r,|\vec{s}-\vec{s}_\bullet|}$$

Count number of possible mutations from r mismatches:

e.g. r=3:

```
TGTGA...TCACA
TATGA...GCTCA
```

$r \to r+1:$	$(L-r)(\mathcal{A}-1)$
$r \to r:$	$r(\mathcal{A}-2)$
$r \to r-1:$	r

$\Rightarrow$ evolution equation in discrete Hamming distance space

$$\begin{aligned}\partial_t N(r,t) &= \Phi(r)N(r,t) + \mu_0\Big[(r+1)N(r+1,t) - rN(r,t)\Big] + \\ &\quad - \mu_0(\mathcal{A}-1)\Big[(L-r)N(r,t) - (L-(r-1))N(r-1,t)\Big]\end{aligned}$$

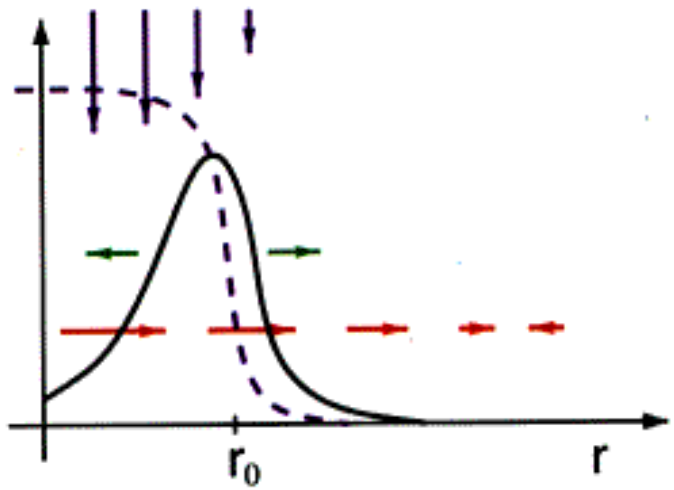
Continuum limit

$\Rightarrow$ Drift-diffusion equation

$$\partial_t n(r,t) = \partial_r \left[D(r) \partial_r n(r,t) - v(r)\, n(r,t) \right] + \left[\phi_0 + \Delta\phi(r) \right] n(r,t)$$

Diffusion coefficient

$$D(r) = \frac{\mu}{2} \left[1 - \frac{\mathcal{A}-2}{\mathcal{A}-1} \frac{r}{L} \right]$$



Drift velocity

$$v(r) = \mu \left[1 - \frac{\mathcal{A}}{\mathcal{A}-1} \frac{r}{L} \right]$$

drives distribution to $\bar{r} = \frac{\mathcal{A}-1}{\mathcal{A}} L$

Source term

$$\Delta\phi(r) = \frac{\phi_0\, \alpha}{1 + e^{\epsilon(r-r_0)}}$$

Selection threshold

Solve continuum equation for the simplified case

$$D(r) \approx D = \mu/2\,, \qquad v(r) \approx v = \mu\,, \qquad \Delta\phi(r) \approx \alpha\,\theta(r_0 - r)$$

$$\Rightarrow \partial_t n(r,t) = D\,\partial_r^2\, n(r,t) - v\,\partial_r\, n(r,t) + [\phi_0 + \Delta\phi(r)]\, n(r,t)$$

(non-hermitian quantum mechanics problem $\rightarrow$ Hatano & Nelson, 1997)

Delocalization transition $\longleftrightarrow$ Eigen's quasispecies transition

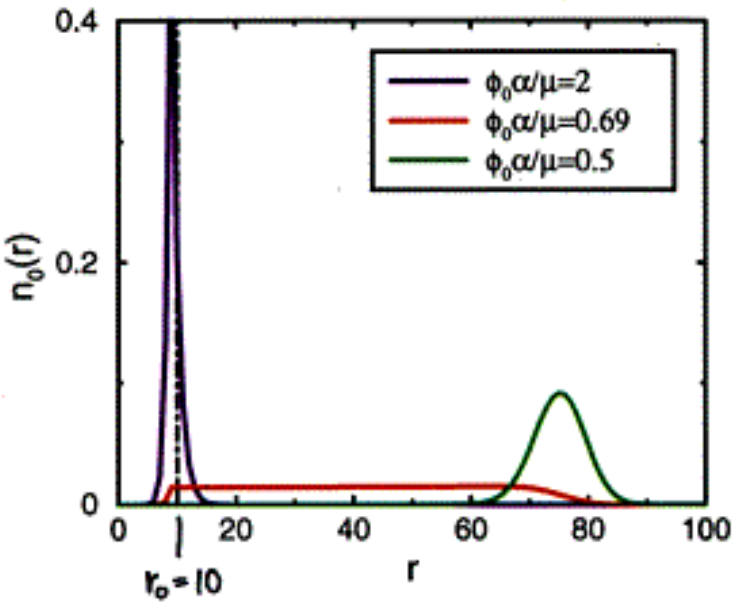
Continuous transition:

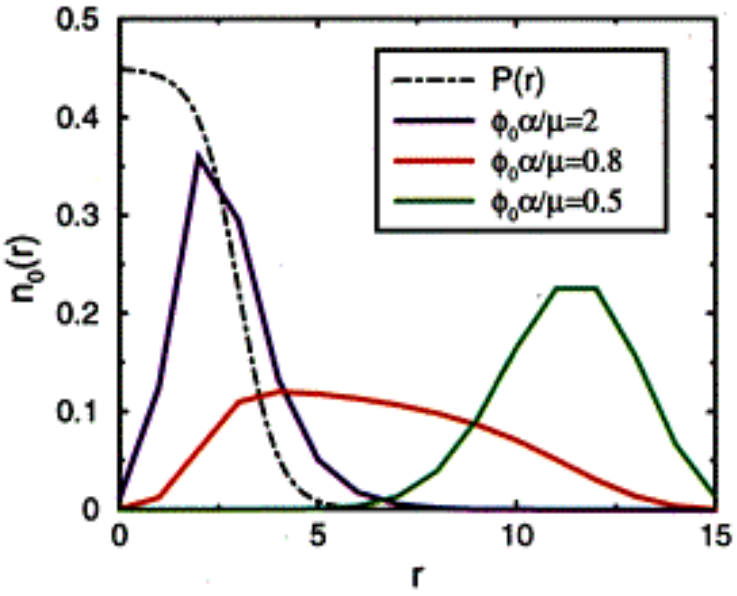
$$\alpha_c = \frac{\mu}{2\phi_0}\left[1 + \frac{\eta_c^2}{r_0^2}\right]\,, \qquad \langle r\rangle \sim (\alpha - \alpha_c)^{-1} \;\text{ for }\; \alpha \gtrsim \alpha_c$$

If α fluctuates in time (due to variations in environment), we expect

$$\langle r\rangle \sim (\alpha - \alpha_c)^{-2} \;\longrightarrow\; \text{(Lubensky \& Nelson, 2000)}$$

Numerical solution





L=100, step function $\sim$ continuum

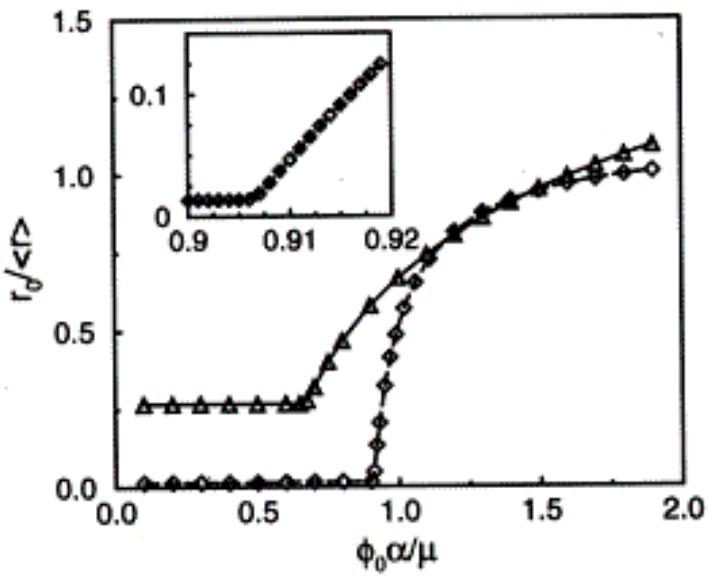
- continuous phase transition
- distribution peak r^* close to r_0

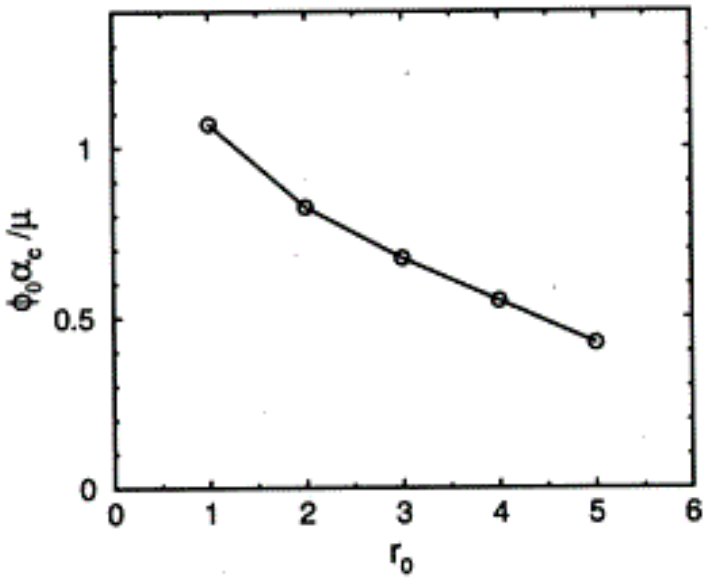
$\Rightarrow$ marginal conservation of motif

L=15, Fermi fct. (ε=2 k_BT, r_0=3)

- still clear transition
- also $r^* \approx r_0$ for $\alpha \gtrsim \alpha_c$
- however $r^* \to 0$ for $\alpha \gg \alpha_c$

Properties of the phase transition

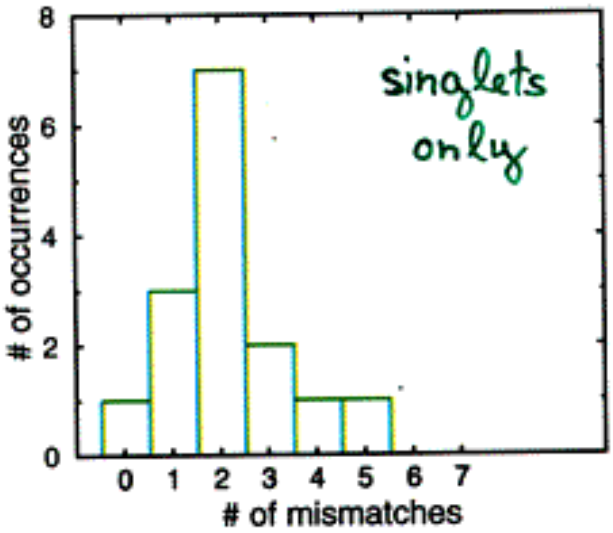


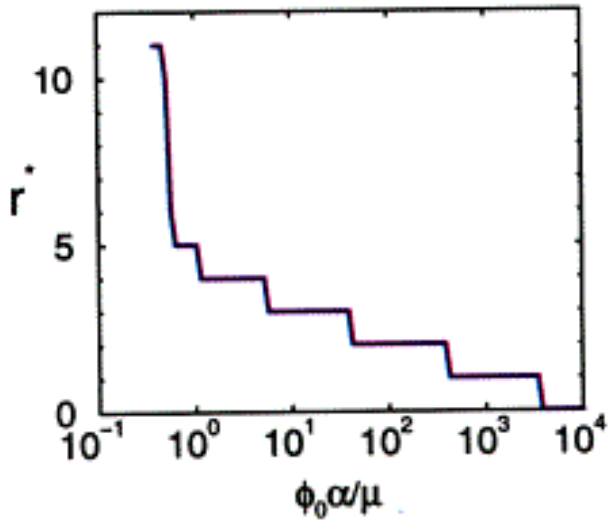


- numerics confirms critical behavior $\langle r\rangle \sim (\alpha - \alpha_c)^{-1}$
- discrete case shows similar behavior
- weak dependence of α_c on r_0
- $\mu \sim 10^{-8}/\tau$, $\phi_0 = \ln 2/\tau \Rightarrow \alpha_c \sim 10^{-8}$, i.e. extremely small

Revisit CRP sites:

....TGTGA.....TCACA....	r
AAAGTGTGACGCCGTGCAAATAAT	2
ATTTCGTGATGTTGCTTGCAAAAA	3
TTTATGTGCGCATCTCCACATTAC	2
ATTCTGTAACAGAGATCACACAAA	1
CCTTTGTGATCGCTTTCACGGAGC	1
AAAACGTGATCAACCCCTCAATTT	3
AACTTGTGGATAAAATCACGGTCT	2
GTTTTGTTACCTGCCTCTAACTTT	3
TTAATTTGAAAATTGGAATATCCA	4
AATTTGCGATGCGTCGCGCATTTT	3
TTAATGAGATTCAGATCACATATA	1
AATGTGTGCGGCAATTCACATTTA	1
GAAACGTGATTTCATGCGTCATTT	5
AAATGACGCATGAAATCACGTTTC	5
TTGCTGTGACTCGATTCACGAAGT	1
TTTTTGTGGCCTGCTTCAAACTTT	2
GAATTGTGACACAGTGCAAATTCA	2
ATAATGTTATACATATCACTCTAA	2
CGATTGTGATTCGATTCACATTTA	0
GTTTTGTGATGGCTATTAGAAATT	2
GAACTGTGAAACGAAACATATTTT	2
AATGTGTGTAAACGTGAACGCAAT	4
TTTGTGTGATCTCTGTTACAGAAT	1
GTAATGTGGAGATGCGCACATAAA	2
TTTTTGCAAGCAACATCACGAAAT	3
TTAATGTGAGTTAGCTCACTCATT	1
ATTATTTGCACGGCGTCACACTTT	2
ATTATTTGAACCAGATCGCATTAC	2
TAATTGTGATGTGTATCGAAGTGT	2





⇒ Can estimate selection pressure on individual binding sequences

What about the multiplets?

Evolution model for doublets

What is the fitness function for a doublet?

$\vec{S}_1$, $\vec{S}_2$ = two binding sequences for the same regulatory protein in promoter region of the same gene

$\Phi_{\vec{S}_1,\vec{S}_2}$ = reproduction rate of the organism in the presence of $\vec{S}_1$ and $\vec{S}_2$

Partial fitness: $\Delta\Phi_{\vec{S}_1,\vec{S}_2} = \Phi_{\vec{S}_1,\vec{S}_2} - \phi_0$

Probability that *at least one* site is occupied by a protein:

$$P_{\vec{S}_1}(1 - P_{\vec{S}_2}) + P_{\vec{S}_2}(1 - P_{\vec{S}_1}) + P_{\vec{S}_1}P_{\vec{S}_2} > P_{\vec{S}_1},\ P_{\vec{S}_2}$$

In general, two proteins may be better than one,

$$\Delta\Phi_{\vec{S}_1,\vec{S}_2} = \alpha\phi_0 \cdot [P_{\vec{S}_1}(1 - P_{\vec{S}_2}) + P_{\vec{S}_2}(1 - P_{\vec{S}_1}) + \omega \cdot P_{\vec{S}_1}P_{\vec{S}_2}]$$

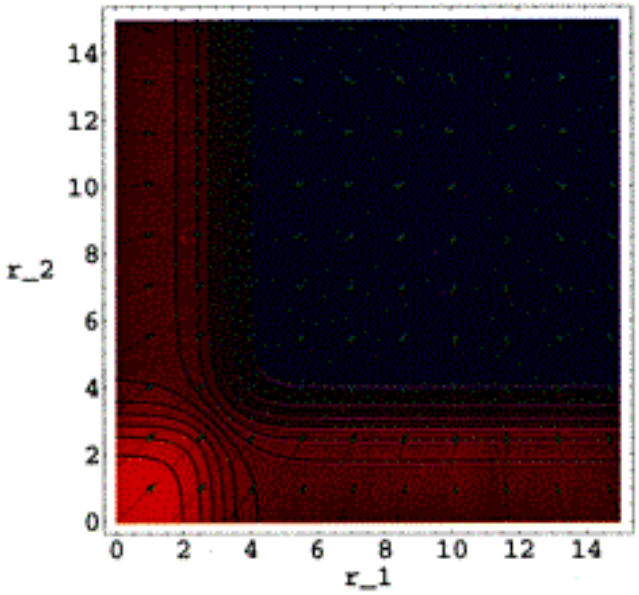
If $\omega > 1$, the two proteins act cooperatively.

2D "radial" evolution equation

r_1 = # mismatches of first site

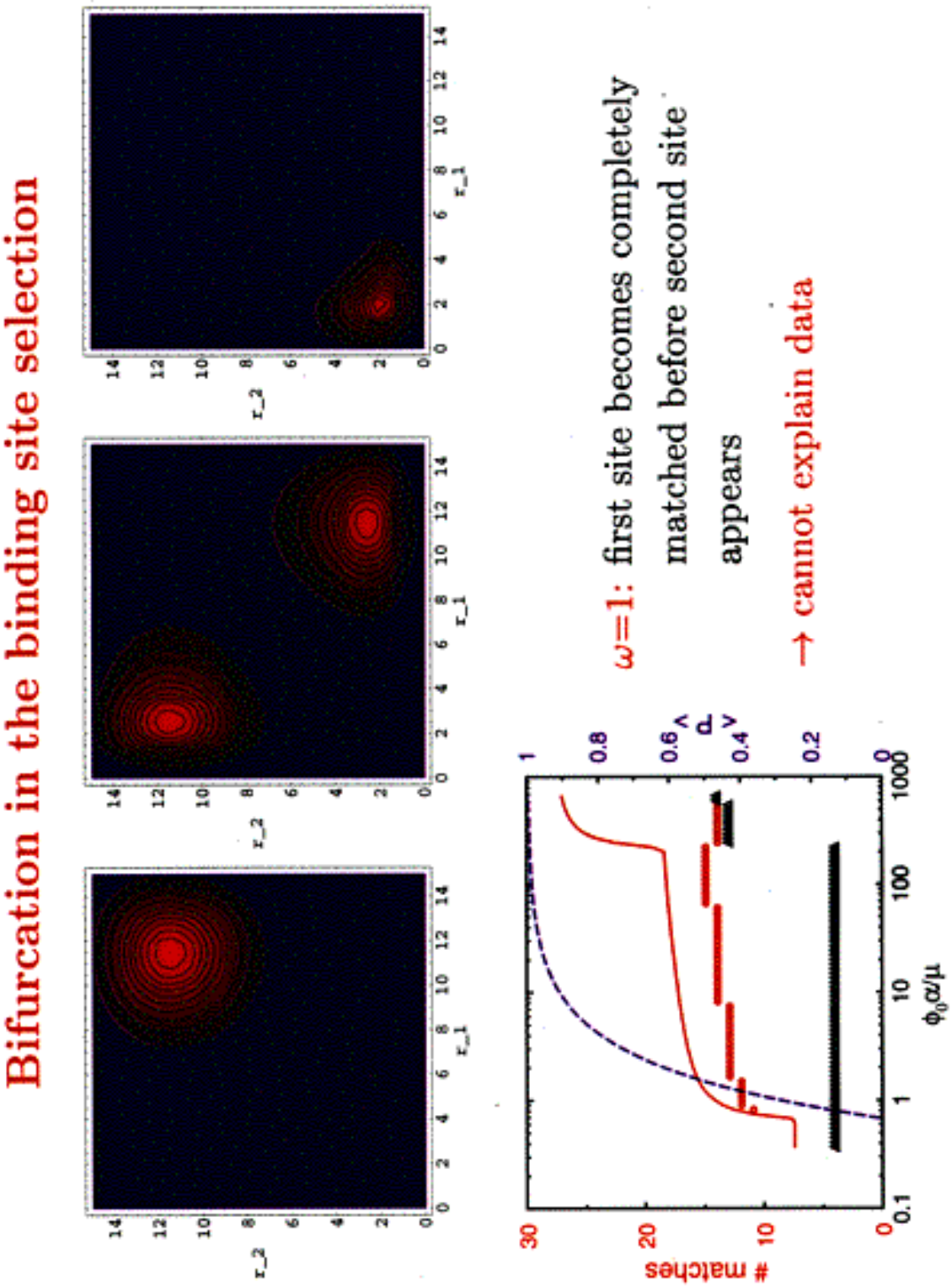
r_2 = # mismatches of second site

"mismatch vector" $\vec{r} = (r_1, r_2)$



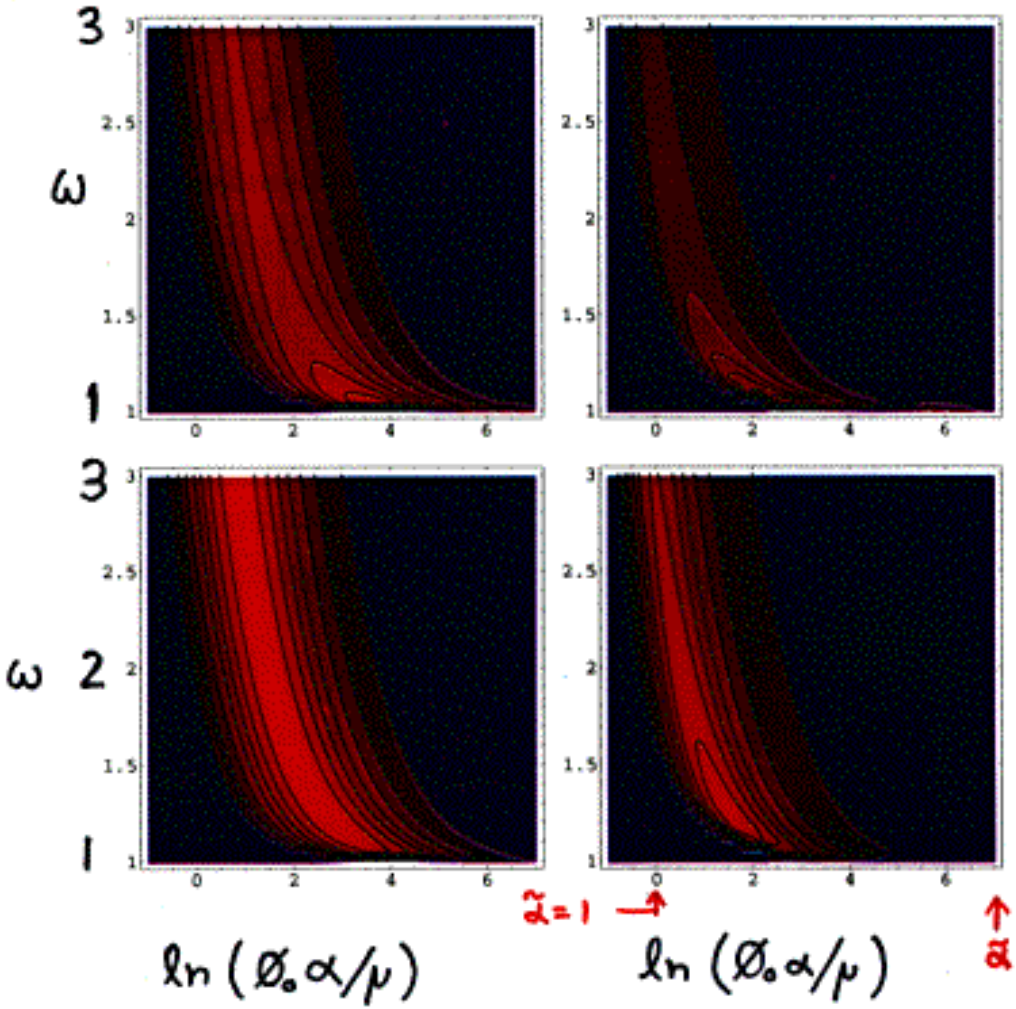
$$\partial_t n(\vec{r},t) = \partial_{r_1}\left[D(r_1)\partial_{r_1} n(\vec{r},t) - v(r_1)\, n(\vec{r},t)\right] + \partial_{r_2}\left[D(r_2)\partial_{r_2} n(\vec{r},t) - v(r_2)\, n(\vec{r},t)\right] + \phi(\vec{r})\, n(\vec{r},t)$$

[15] Dr. U. Gerland, UCSD (ITP 4/25/01) Discussion: Selection of DNA Motifs via Protein Binding: Survival



Estimation of α and ω for individual sites

Strategy: plot normalized $n_0(r_1, r_2; \alpha, \omega)$ at fixed (r_1, r_2) as a function of (α, ω)



Index:

(1,3) (1,5)

(2,2) (2,4)

$\Rightarrow \omega \approx 1.3 \pm 0.2$ required for explanation of (2,4) site