Sequence evolution in phenotypic space: An overview

Luca Peliti

Dipartimento di Scienze Fisiche, Unità CNISM and Sezione INFN

Università “Federico II”, Napoli, Italy

Summary

- Nucleotide sequences are not *direct* targets of molecular evolution, but are selected depending on the *traits* they determine

Summary

- Nucleotide sequences are not *direct* targets of molecular evolution, but are selected depending on the *traits* they determine
- The sequence $\longrightarrow$trait application can in some cases be regarded as a “smooth” one, similar to microstate$\longrightarrow$observable application in Statistical Mechanics

Summary

- Nucleotide sequences are not *direct* targets of molecular evolution, but are selected depending on the *traits* they determine
- The sequence $\longrightarrow$trait application can in some cases be regarded as a “smooth” one, similar to microstate$\longrightarrow$observable application in Statistical Mechanics
- I shall look at the “equilibrium” aspects of the resulting ensembles rather than at the process of evolving adaptation (Krug, Fisher-Desai)

Summary

- Nucleotide sequences are not *direct* targets of molecular evolution, but are selected depending on the *traits* they determine
- The sequence $\longrightarrow$trait application can in some cases be regarded as a “smooth” one, similar to microstate$\longrightarrow$observable application in Statistical Mechanics
- I shall look at the “equilibrium” aspects of the resulting ensembles rather than at the process of evolving adaptation (Krug, Fisher-Desai)
- Contrast between quasispecies theory (long sequence, few traits, very large population) and Kimura-Ohta ensemble (short sequences, finite population)

Summary

- Nucleotide sequences are not *direct* targets of molecular evolution, but are selected depending on the *traits* they determine
- The sequence $\longrightarrow$trait application can in some cases be regarded as a “smooth” one, similar to microstate$\longrightarrow$observable application in Statistical Mechanics
- I shall look at the “equilibrium” aspects of the resulting ensembles rather than at the process of evolving adaptation (Krug, Fisher-Desai)
- Contrast between quasispecies theory (long sequence, few traits, very large population) and Kimura-Ohta ensemble (short sequences, finite population)
- Empirical checks (Mustonen, Lässig, . . .)

A description

Genotype: Sequence of units (e.g. nucleotides)

$$\mathbf{g} = (\mathrm{g}_i); \quad \mathrm{g}_i \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}; \qquad i = 1, \ldots, L$$

A description

Genotype: Sequence of units (e.g. nucleotides)

$$\mathbf{g} = (\mathrm{g}_i); \quad \mathrm{g}_i \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}; \qquad i = 1, \ldots, L$$

In general: $\mathrm{g}_i \in \{1, \ldots, c\}$

A description

Genotype: Sequence of units (e.g. nucleotides)

$$\mathbf{g} = (\mathrm{g}_i); \quad \mathrm{g}_i \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}; \qquad i = 1, \ldots, L$$

In general: $\mathrm{g}_i \in \{1, \ldots, c\}$

Phenotype: Collection of smooth "observables"

$$\vec{m} = (m^\alpha); \quad m^\alpha(\mathbf{g}) = \frac{1}{L} \sum_i \xi^\alpha_{i\mathrm{g}_i}; \quad \alpha = 1, \ldots, r$$

(Linear model)

Mutation-selection (Quasispecies) model

- Asexual reproduction of “genotypes” $\{\mathbf{g}\}$

Mutation-selection (Quasispecies) model

- Asexual reproduction of “genotypes” $\{\mathbf{g}\}$
- Reproduction rate (“fitness”) $LF\left(\vec{m}(\mathbf{g})\right)$: proportional to length L ($1/L$ is a smoothness parameter)

Mutation-selection (Quasispecies) model

- Asexual reproduction of “genotypes” $\{\mathbf{g}\}$
- Reproduction rate (“fitness”) $LF\left(\vec{m}(\mathbf{g})\right)$: proportional to length L ($1/L$ is a smoothness parameter)
- Point-wise mutations: $\mathrm{g}_i \overset{\mu}{\rightleftharpoons} \mathrm{g}_i'$

Mutation-selection (Quasispecies) model

- Asexual reproduction of “genotypes” $\{\mathbf{g}\}$
- Reproduction rate (“fitness”) $LF\left(\vec{m}(\mathbf{g})\right)$: proportional to length L ($1/L$ is a smoothness parameter)
- Point-wise mutations: $\mathrm{g}_i \overset{\mu}{\rightleftharpoons} \mathrm{g}_i'$
- Infinite population limit $\Longrightarrow$ deterministic process

Mutation-selection (Quasispecies) model

- Asexual reproduction of “genotypes” $\{\mathbf{g}\}$
- Reproduction rate (“fitness”) $LF\left(\vec{m}(\mathbf{g})\right)$: proportional to length L ($1/L$ is a smoothness parameter)
- Point-wise mutations: $\mathrm{g}_i \overset{\mu}{\rightleftharpoons} \mathrm{g}_i'$
- Infinite population limit $\Longrightarrow$ deterministic process
- **Para**llel **mu**tation-**se**lection mechanism

The Quasispecies equation

$n(\mathbf{g}, t)$: fraction of individuals carrying genotype $\mathbf{g}$
$(\sum_{\mathbf{g}} n(\mathbf{g}, t) = 1, \ \forall t)$

Eigen (1971)

The Quasispecies equation

$n(\mathbf{g}, t)$: fraction of individuals carrying genotype $\mathbf{g}$
($\sum_{\mathbf{g}} n(\mathbf{g}, t) = 1, \ \forall t$)

Eigen (1971)

$$\frac{\partial}{\partial t} n(\mathbf{g}, t) \quad =$$

The Quasispecies equation

$n(\mathbf{g}, t)$: fraction of individuals carrying genotype $\mathbf{g}$
($\sum_{\mathbf{g}} n(\mathbf{g}, t) = 1, \ \forall t$)

Eigen (1971)

$$\frac{\partial}{\partial t} n(\mathbf{g}, t) \quad = \quad \underbrace{\sum_{\mathbf{g}'} \left[M_{\mathbf{g}\mathbf{g}'} n(\mathbf{g}', t) - M_{\mathbf{g}'\mathbf{g}} n(\mathbf{g}, t) \right]}_{\text{mutation}}$$

The Quasispecies equation

$n(\mathbf{g}, t)$: fraction of individuals carrying genotype $\mathbf{g}$
($\sum_{\mathbf{g}} n(\mathbf{g}, t) = 1, \ \forall t$)

Eigen (1971)

$$\frac{\partial}{\partial t} n(\mathbf{g}, t) = \underbrace{\sum_{\mathbf{g}'} \left[M_{\mathbf{g}\mathbf{g}'} n(\mathbf{g}', t) - M_{\mathbf{g}'\mathbf{g}} n(\mathbf{g}, t) \right]}_{\text{mutation}} + \underbrace{\left(LF(\vec{m}(\mathbf{g})) - \Lambda(\{n\}) \right) n(\mathbf{g}, t)}_{\text{selection}}$$

The Quasispecies equation

$n(\mathbf{g}, t)$: fraction of individuals carrying genotype $\mathbf{g}$
($\sum_{\mathbf{g}} n(\mathbf{g}, t) = 1, \ \forall t$)

Eigen (1971)

$$\frac{\partial}{\partial t} n(\mathbf{g}, t) = \underbrace{\sum_{\mathbf{g}'} \left[M_{\mathbf{g}\mathbf{g}'} n(\mathbf{g}', t) - M_{\mathbf{g}'\mathbf{g}} n(\mathbf{g}, t) \right]}_{\text{mutation}}$$

$$+ \underbrace{\left(LF(\vec{m}(\mathbf{g})) - \Lambda(\{n\}) \right) n(\mathbf{g}, t)}_{\text{selection}}$$

$$\Lambda(\{n\}) = \sum_{\mathbf{g}} LF(\vec{m}(\mathbf{g})) n(\mathbf{g}, t) \qquad \text{(normalization)}$$

Evolution equation for the genotypes

"Slowly varying" fitness ($L \gg 1$):

$$\lambda_{i\mathbf{g}_i} = \sum_\alpha \frac{\partial F}{\partial m^\alpha} \xi^\alpha_{i\mathbf{g}_i}$$

Factorization:

$$n(\mathbf{g}, t) = \prod_i n_i(\mathbf{g}_i, t)$$

Equation for $n_i(\mathrm{g}_i, t)$:

$$\partial_t n_i(\mathrm{g}, t) = \left(\frac{\mu}{c-1}\right) \sum_{\mathrm{g}' \ (\neq \mathrm{g})} \left[n_i(\mathrm{g}', t) - n_i(\mathrm{g}, t)\right] + \left(\lambda_{i\mathrm{g}} - \langle\lambda\rangle\right) n_i(\mathrm{g}, t)$$

where $\langle\lambda\rangle = \sum_{\mathrm{g}} \lambda_{i\mathrm{g}} n(\mathrm{g}, t)$

N.B.: $\sum_{\mathrm{g}' \ (\neq \mathrm{g})} n_i(\mathrm{g}', t) = 1 - n_i(\mathrm{g}, t)$

Evolution equation for the phenotypes

$$\langle m^\alpha \rangle = \sum_{\mathbf{g}} m^\alpha(\mathbf{g})\, n(\mathbf{g}, t); \quad \Delta^{\alpha\beta} = \left\langle m^\alpha m^\beta \right\rangle - \langle m^\alpha \rangle \left\langle m^\beta \right\rangle$$

$$\partial_t \langle m^\alpha \rangle = -\left(\frac{\mu c}{c-1}\right) \langle m^\alpha \rangle + \sum_i \sum_{\mathrm{g}_i} \xi^\alpha_{i\mathrm{g}_i} \lambda_{i\mathrm{g}_i} n_i(\mathrm{g}_i, t)$$

$$- \sum_i \left(\sum_{\mathrm{g}'_i} \xi^\alpha_{i\mathrm{g}'_i} \lambda_{i\mathrm{g}'_i} n_i(\mathrm{g}'_i, t) \right) \left(\sum_{\mathrm{g}_i} \xi^\alpha_{i\mathrm{g}_i} n_i(\mathrm{g}_i, t) \right)$$

Evolution equation for the phenotypes

$$\langle m^\alpha \rangle = \sum_{\mathbf{g}} m^\alpha(\mathbf{g})\, n(\mathbf{g}, t); \quad \Delta^{\alpha\beta} = \left\langle m^\alpha m^\beta \right\rangle - \langle m^\alpha \rangle \left\langle m^\beta \right\rangle$$

$$\begin{aligned}
\partial_t \langle m^\alpha \rangle &= -\left(\frac{\mu c}{c-1}\right) \langle m^\alpha \rangle + \sum_i \sum_{\mathrm{g}_i} \xi^\alpha_{i\mathrm{g}_i} \lambda_{i\mathrm{g}_i} n_i(\mathrm{g}_i, t) \\
&\quad - \sum_i \left(\sum_{\mathrm{g}'_i} \xi^\alpha_{i\mathrm{g}'_i} \lambda_{i\mathrm{g}'_i} n_i(\mathrm{g}'_i, t) \right) \left(\sum_{\mathrm{g}_i} \xi^\alpha_{i\mathrm{g}_i} n_i(\mathrm{g}_i, t) \right) \\
&= \underbrace{-\left(\frac{\mu c}{c-1}\right) \langle m^\alpha \rangle}_{\text{Mutations}} + \underbrace{\sum_\beta \Delta^{\alpha\beta} \left.\frac{\partial F}{\partial m^\beta}\right|_{\vec{m}=\langle \vec{m} \rangle}}_{\text{Fisher's optimization}}
\end{aligned}$$

What is optimized in evolution?

Let us define

$$\mathcal{F}(\vec{h}) = \ln \sum_{\mathbf{g}} \exp \left\{ \sum_{\alpha} h^{\alpha} m^{\alpha}(\mathbf{g}) \right\}$$

Then

$$\begin{aligned} \left\langle m^{\alpha}(\vec{h}) \right\rangle &= \frac{\partial \mathcal{F}}{\partial h^{\alpha}}, \quad \alpha = 1, \ldots, r \\ \mathcal{S}(\vec{m}) &= \mathcal{F}\left(\vec{h}(\vec{m})\right) \end{aligned}$$

A Lyapunov function

Thus the quantity

$$\boxed{\Gamma(\vec{m}) = F(\vec{m}) - \frac{\mu c}{c-1}\mathcal{S}(\vec{m})}$$

satisfies

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}t} = \sum_{\alpha\beta} \Delta^{-1}_{\alpha\beta} \left\langle \dot{m}^{\alpha} \right\rangle \left\langle \dot{m}^{\beta} \right\rangle \geq 0$$

The “mesa” landscape

Threshold landscape on a trait
$F(\vec{m})$: “fitness landscape”

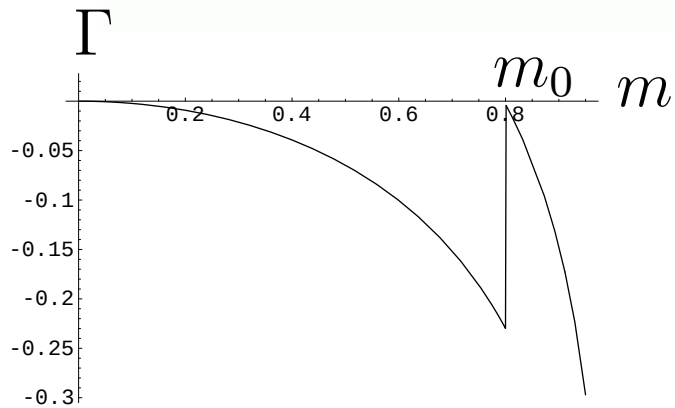
$$F(m) = \lambda\,\theta(m - m_0)$$

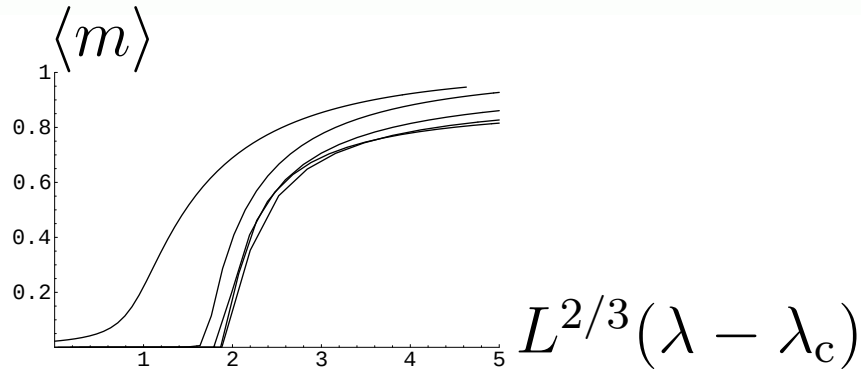
$$r = 1,\, c = 2 \Rightarrow \mathrm{g}_i = \pm 1$$

$$\begin{aligned} m(\mathbf{g}) &= L^{-1} \sum_i \mathrm{g}_i \\ \Gamma &= F(m) - \frac{\mu}{2} \ln\left(1 - m^2\right) \end{aligned}$$

$$\langle m \rangle = \begin{cases} 0, & \text{for } \lambda < \lambda_{\mathrm{c}} = \mu \ln(1 - m_0^2); \\ m_0, & \text{for } \lambda > \lambda_{\mathrm{c}} \end{cases}$$

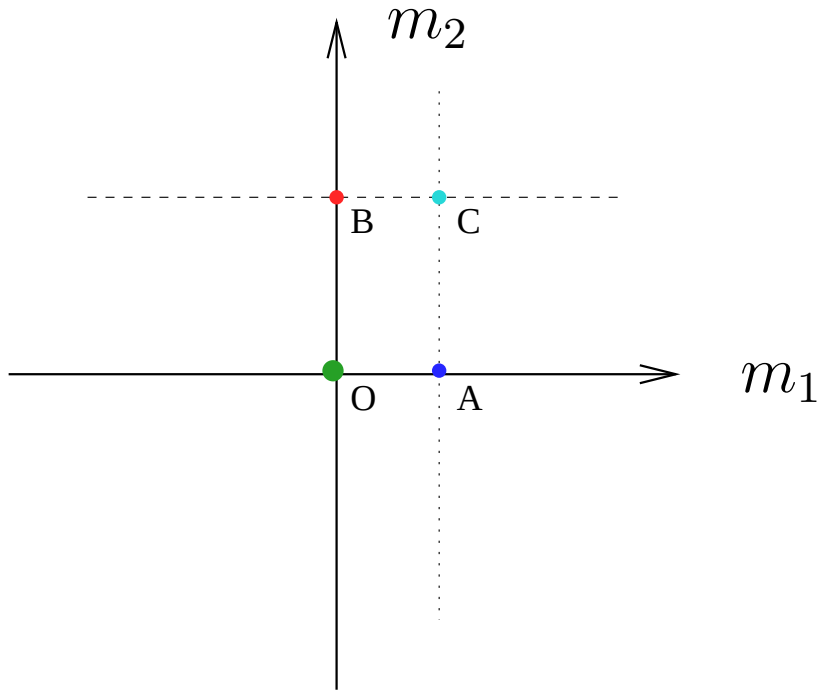
Phase transition





"Mesa" landscape: Data for $L = 10, 20, 40, 80, 160$

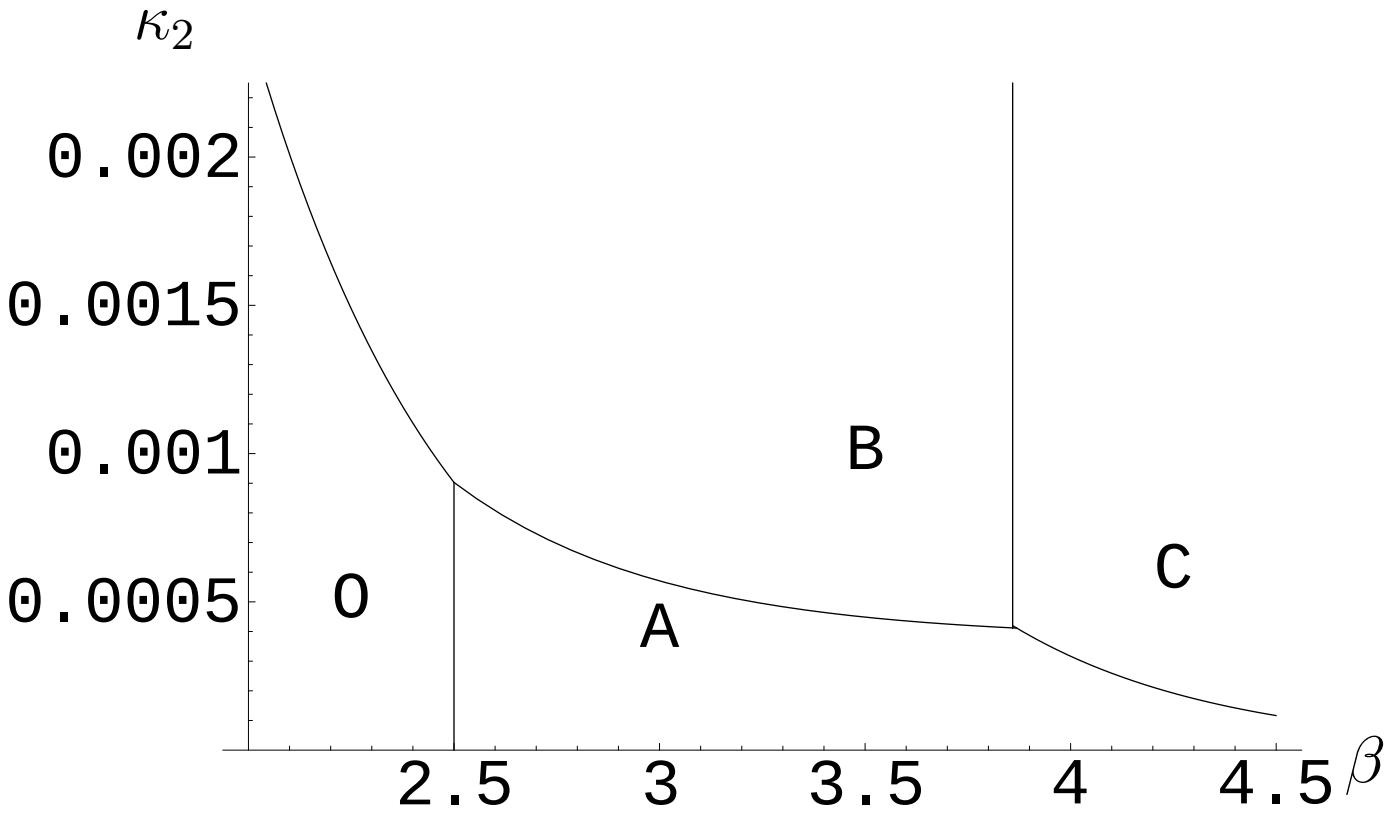
The 2D Hopfield landscape



$$m^\alpha = \frac{1}{L}\sum_i \xi_i^\alpha \qquad i = 1, 2$$

$$F(\vec{m}) = \kappa_1 \theta(m^1 - 0.33) + \kappa_2 \theta(m^2 - 0.5)$$

Phase diagram

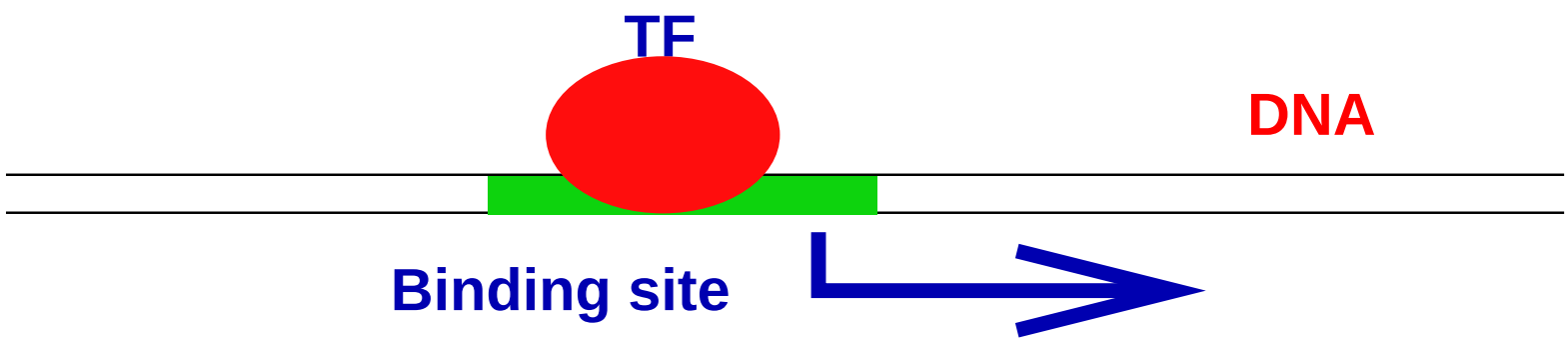


Stability regions for the points O: $\vec{m}^* = (0, 0)$, A: $\vec{m}^* = (0.33, 0)$, B: $\vec{m}^* = (0, 0.5)$, and C: $\vec{m}^* = (0.33, 0.5)$

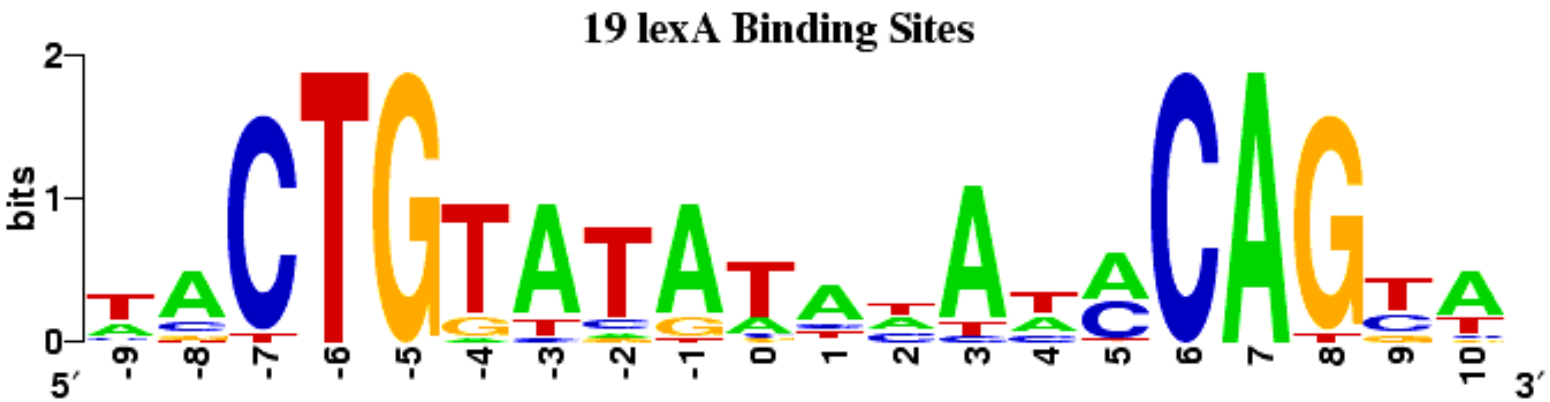
Observations

- The equilibrium point is at the border of the selected region
- Trait mutations are (largely) deleterious, most sequence mutations are neutral
- Linear genome $\longrightarrow$ trait mapping, yet nonlinear fitness effects
- Analysis under the assumptions of long sequences and very large populations

Transcription factors



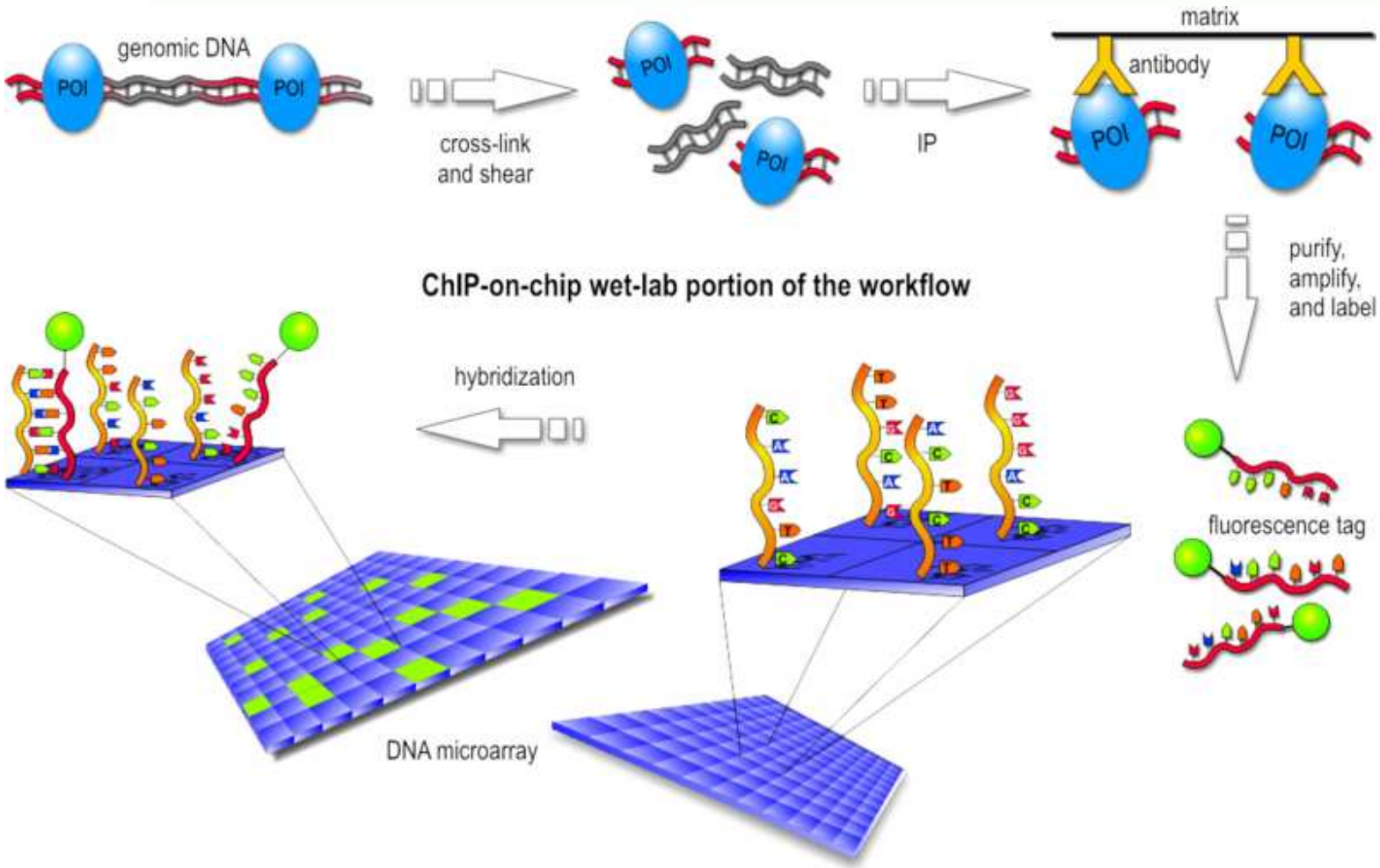
Sequence conservation logo:



TF binding sites properties

- Binding sites are relatively short ($10 \div 20$ sites)
- Binding sites for the *same* factor can have quite different sequences along the genome
- Binding energies can be evaluated “in vitro”
- High-throughput data on protein-DNA binding are becoming available (“ChIP on chip”)

ChIP on chip


genomic DNA
POI
POI
cross-link
and shear
POI
POI
IP
matrix
antibody
POI
POI
purify,
amplify,
and label
ChIP-on-chip wet-lab portion of the workflow
hybridization
fluorescence tag
DNA microarray

Linear model of binding

$$E(\mathbf{g}) = \sum_{i=1}^{L} \epsilon_i(\mathrm{g}_i)$$

Heuristic evaluation of $\epsilon_i(\mathrm{g})$:

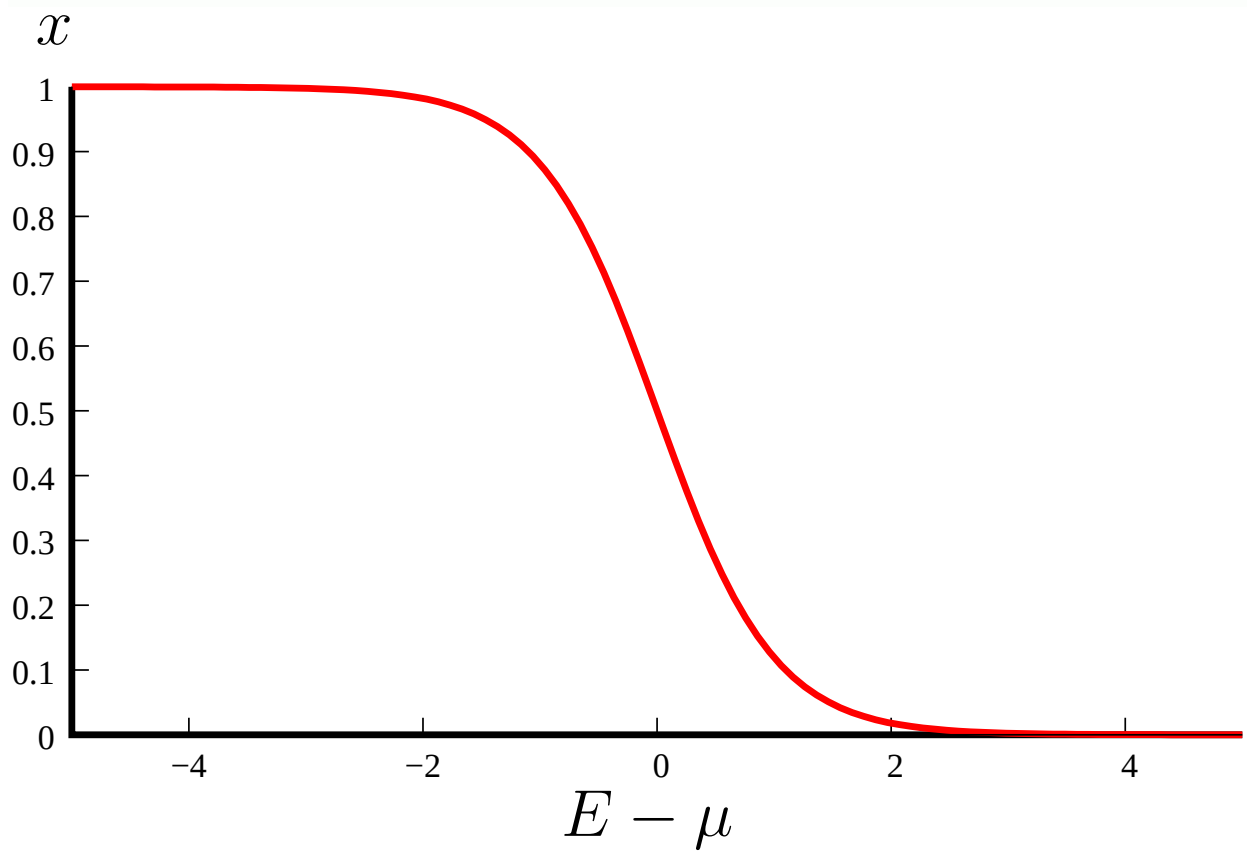
- $q_i(\mathrm{g})$: frequency of nucleotide g at site i in validated binding sites
- $p_0(\mathrm{g}))$: overall frequency of nucleotide g at site i in the intergenic sequence
- Assuming $q_i(\mathrm{g}) \propto \mathrm{e}^{-\beta\epsilon_i(\mathrm{g})}$

$$\epsilon_i(\mathrm{g}) = \beta^{-1} \log \frac{q_i(\mathrm{g})}{p_0(\mathrm{g})}$$

Berg and von Hippel, 1987; Stormo and Fields, 1998

More reliable estimates on the parameters $\epsilon_i(\mathrm{g})$ can be obtained from ChIP on chip data: cf. *Kinney, Tkačik and Callan*)

Binding probability



N.B. The TF chemical potential μ can depend on the state of the cell cycle

Considerations and assumptions

- Assumption: the fitness is a function of the binding energy

Considerations and assumptions

- Assumption: the fitness is a function of the binding energy
- This function is in general highly nonlinear ($\propto x$ or more complicated)

Considerations and assumptions

- Assumption: the fitness is a function of the binding energy
- This function is in general highly nonlinear ($\propto x$ or more complicated)
- Low mutation rate per nucleotide & short length of a binding site $\longrightarrow$ few mutants of a given binding site are present in the population

Considerations and assumptions

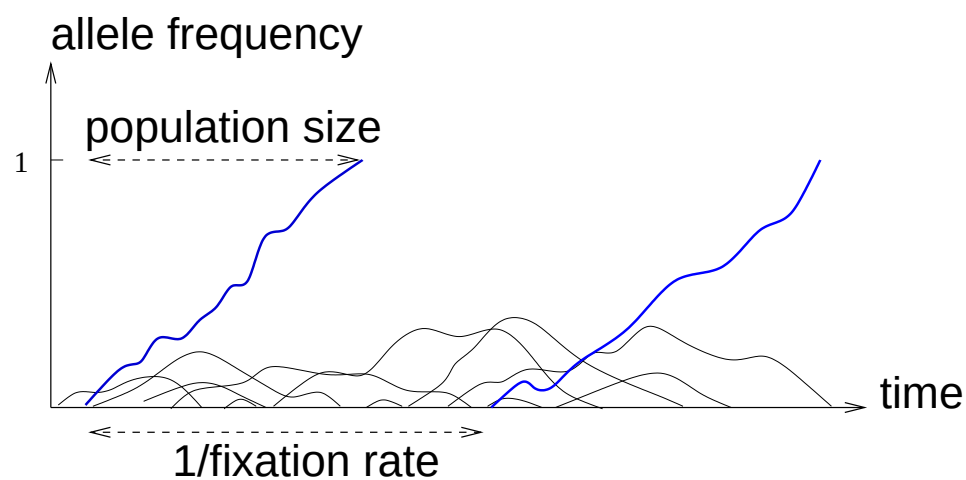
- Assumption: the fitness is a function of the binding energy
- This function is in general highly nonlinear ($\propto x$ or more complicated)
- Low mutation rate per nucleotide & short length of a binding site $\longrightarrow$ few mutants of a given binding site are present in the population
- Fluctuation-based (Kimura-Ohta) evolution model

The Kimura-Ohta process

- Rare mutations: comparison of a mutant against a homogeneous population
- Fitness change:

$$\Delta F = F(\text{mutant}) - F(\text{wild type})$$

- Fraction $x(t)$ of mutants: biased random walk



The fixation probability

- Fixation probability:

$$p^{\mathrm{fix}} = \frac{1 - \mathrm{e}^{-2\,\Delta F}}{1 - \mathrm{e}^{-2N\,\Delta F}}$$

- Limiting behaviors:
  - $\Delta F = 0 \Rightarrow p^{\mathrm{fix}} = 1/N$ (neutral theory)
  - $\Delta F \simeq -(1/N) \Rightarrow p^{\mathrm{fix}} \to 0$
  - $\Delta F \simeq (1/N) \Rightarrow p^{\mathrm{fix}} \to 1$
- Selection intensity range $\sim (1/N)$

Mutation and selection

- Mutation $\mathbf{g} \longrightarrow \mathbf{g}'$ with probability $u_{\mathbf{g}'\,\mathbf{g}}$
- Mutation "detailed balance" $\Rightarrow$ background distribution $P_0(\mathbf{g})$

$$u_{\mathbf{g}'\,\mathbf{g}} P_0(\mathbf{g}) = u_{\mathbf{g}\,\mathbf{g}'} P_0(\mathbf{g}')$$

- Stationary distribution under selection:

$$\boxed{Q(\mathbf{g}) = P_0(\mathbf{g})\, \mathrm{e}^{2NF(\mathbf{g})}}$$

Berg, Willman and Lässig, 2004

Interpretation

- $Q(\mathbf{g})$ appears as the *distribution in time* of the sequences $\mathbf{g}$
- **Ergodic hypothesis:** the distribution of functional sequences *along the genome* is a sample from $Q(\mathbf{g})$
- The background distribution $P_0(\mathbf{g})$ corresponds to the distribution of *nonfunctional* sequences
- The fitness function $F(E)$ can be inferred from the *actual* distribution of binding energies in functional and nonfunctional sequences:

$$2N\,F(E) = \log \frac{Q(E)}{P_0(E)}$$

Sequence distribution

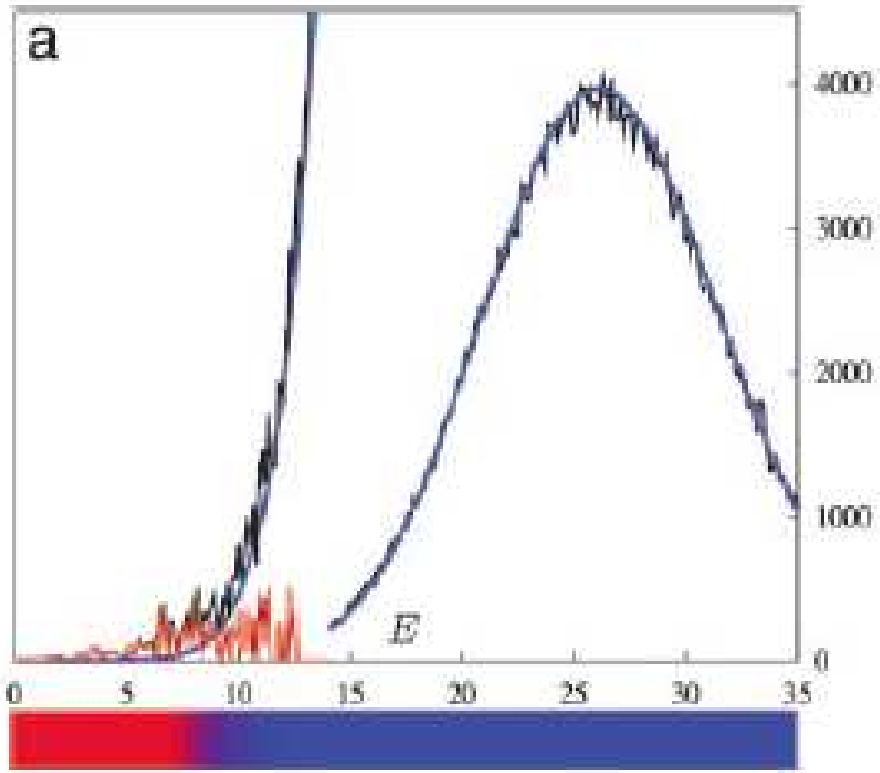
- The binding-energy distribution $W(E)$ along the genome is the superposition of the functional distribution $Q(E)$ and of the background distribution $P_0(E)$:

$$W(E) = (1-\lambda)P_0(E) + \lambda Q(E)$$

  λ: fraction of functional sites

- Problem: $Q(E)$ and $P_0(E)$ overlap

Binding energy distribution for CRP in E. coli



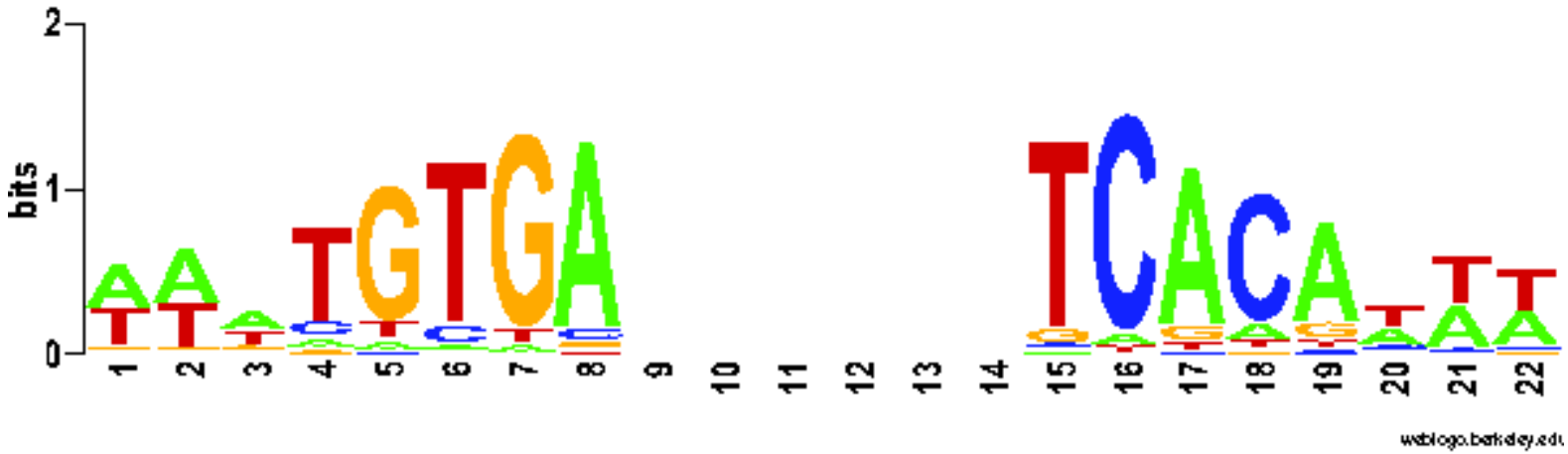
CRP: cAMP response protein

Energy units: arbitrary Left: 30-fold zoom in the region $E < 14$

Color code: probability of being functional according to the model

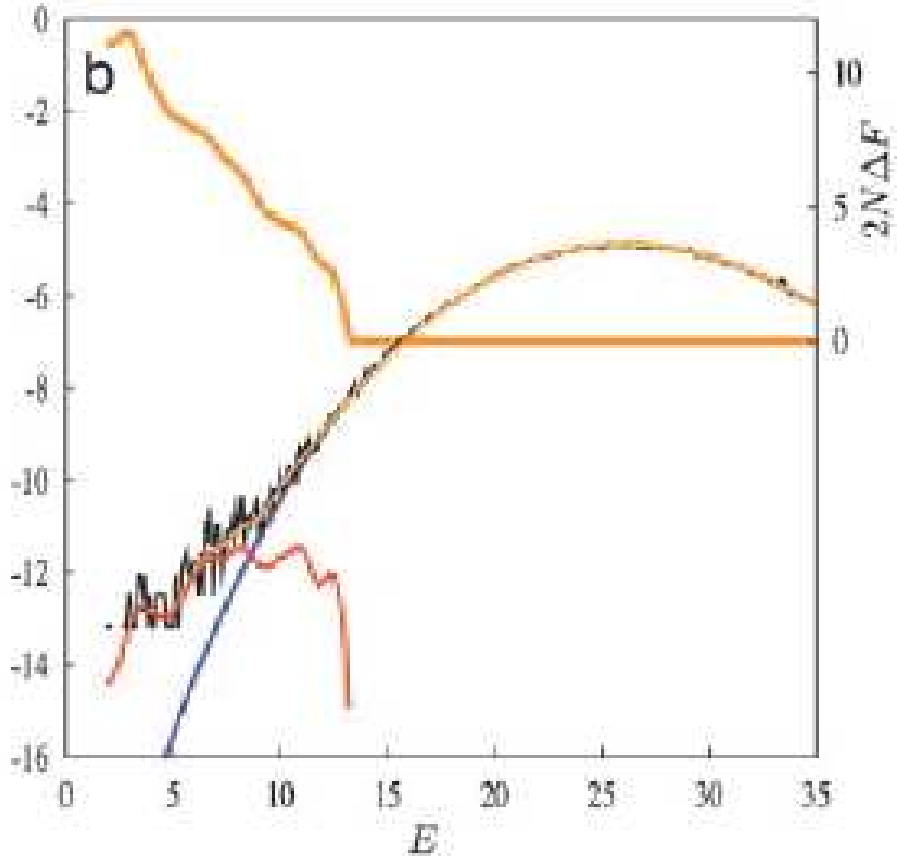
Mustonen and Lässig, 2004

CRP logo



Carmack et al., 2007

The inferred $F(E)$



Same distribution, log on the y-axis

Red: excess count wrt background

Orange: inferred $2N\,\Delta F(E) = \log\left[Q(E)/P_0(E)\right]$

Cross-species comparison

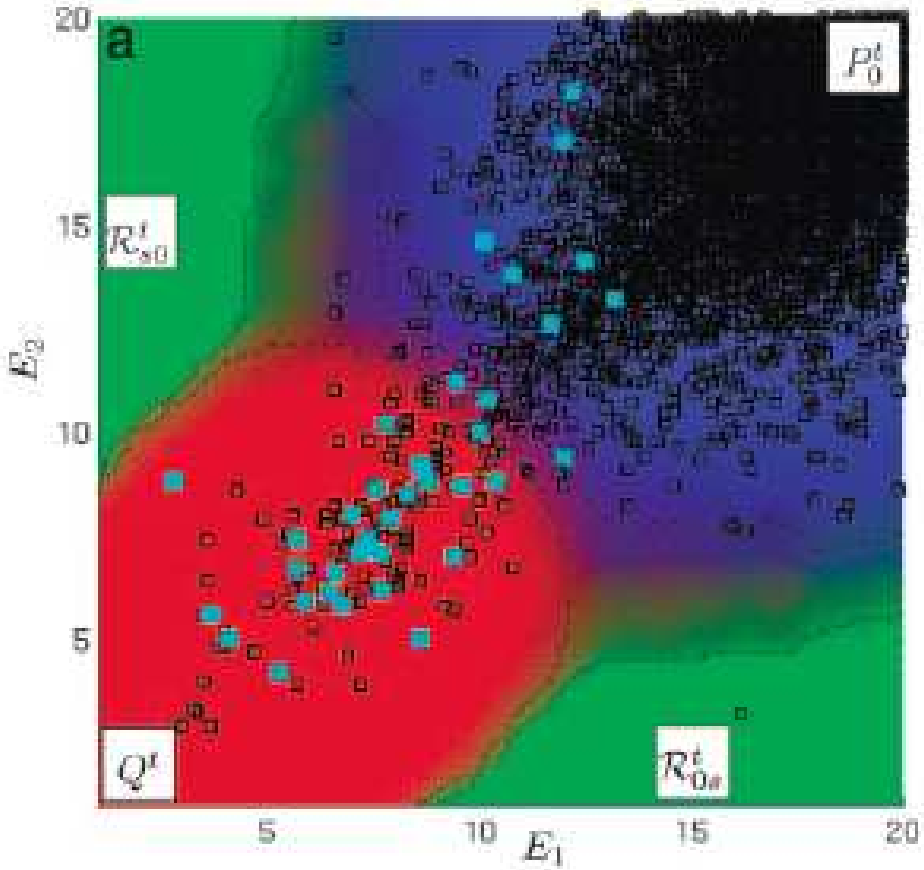
- Binding energy distribution of orthologous sites across species
- Joint pdf after a divergence time t

$$
\begin{aligned}
P_0^t(E_1, E_2) &= G_0^t(E_2|E_1)P_0(E_1) && \text{nonfunctional} \\
Q^t(E_1, E_2) &= G_s^t(E_2|E_1)Q(E_1) && \text{functional} \\
W(E_1, E_2) &= (1-\lambda)P_0^t(E_1, E_2) + \lambda Q^t(E_1, E_2)
\end{aligned}
$$

  G's are numerically evaluated from the model

- Generalization to gain and loss of function
- Comparison with pair energy distribution

E. coli vs. S. typhimurium



Color code: likelihood of functional history: blue (conserved neutrality), red (conserved function), green (functional switch)

Light blue dots: verified binding sites in *E. coli*

Loss and gain of function

- The distributions $P_0(E)$ and $Q(E)$ overlap
- This is a nuisance for us, but can be an evolutionary advantage
- This depends on the value of μ
- Setting μ such that there is a large enough “reservoir” of nonfunctional random sequences can enhance evolvability
- Small changes of sequences can lead to loss and gain of function
- Sequences which have become functional from the reservoir have a small likelihood of switching, but there are may potential candidates

Comments

- The model predicts a binding energy distribution which nicely compares with the observed one
- Increased bonus by comparison of *phenotype* distribution of orthologous sequences in more than two species
- More recent results, based on an *independent* estimation of the binding energy ChIP on chip), confirm the approach in *S. cervisiae* (*Mustonen et al., unpublished*)
- Evidence for the coexistence of functions: overlapping and non-overlapping sequences (analogous to the $2D$-Hopfield landscape)

Epistatic effects

- Within the linear model

$$E(\mathbf{g}) = \sum_i \epsilon_i(\mathrm{g}_i)$$

- Thus, if substitutions at different locations are independent, between orthologous sequences

$$\left\langle \Delta E^2 \right\rangle = \underbrace{\sum_i \left\langle \Delta \epsilon_i^2 \right\rangle}_{\delta^2} \propto \text{divergence time}$$

- Cross-species comparison: $\langle \Delta E^2 \rangle$ grows more slowly than δ^2: There must be *compensatory mutations*

Conclusions and outlook

- It has been possible to analyze genome-wide the two-layered sequence $\longrightarrow$ phenotype $\longrightarrow$ fitness model
- This builds on the availability of a large sequence sample from a given organism
- So far, DNA-binding proteins are the premier choice for these studies
- Can we identify other phenotypic traits that can be similarly studied?

Bibliography

M. Eigen, *Naturwissenschaften* **58**, 465 (1971)

L. Peliti, *Europhys. Lett.* **57**, 745 (2002); M. Lässig, L. Peliti, F. Tria, *Europhys. Lett.* **58**, 465 (2003)

V. Mustonen, M. Lässig, *PNAS* **102**, 15936 (2005)

J.B. Kinney, G. Tkačik, C.G. Callan Jr, *PNAS* **104**, 501 (2007)

O.G. Berg, P.H. von Hippel, *J. Mol. Biol.* **193**, 723 (1987)

G. Stormo, D.S. Fields, *Trends BioChem. Sci.* **23**, 109 (1998)

M. Kimura, T. Ohta, *Genetics* **61**, 763 (1969)

J. Berg, S. Willman, M. Lässig, *BMC Eovl. Biol.* **4**, 42 (2004)

V. Mustonen, J. Kinney, C.G. Callan, M. Lässig, to appear