Evolutionary modes of regulatory sequences in eukaryotes

Armita Nourmohammad

Institute of Theoretical Physics
University of Cologne

February 10, 2011

General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation

by Jacques Monod and François Jacob

Services de Biochimie Cellulaire et de Génétique Microbienne, Institut Pasteur, Paris

One conclusion which was repeatedly emphasized is the wide-spread occurrence and the extreme importance of regulatory mechanisms in cellular physiology.

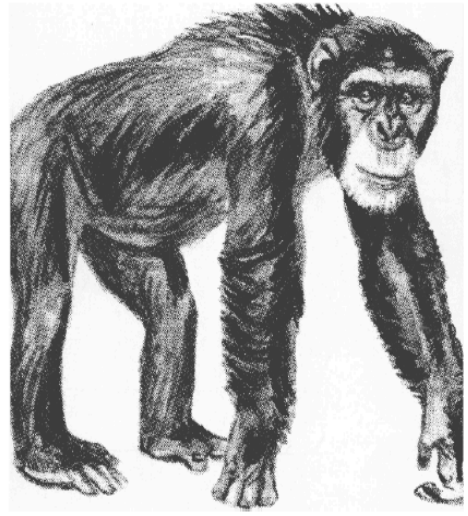
Evolution at Two Levels in Humans and Chimpanzees

Their macromolecules are so alike that regulatory mutations may account for their biological differences.

Mary-Claire King and A. C. Wilson

SCIENCE 11 April 1975
vol. 188, No. 4184
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Soon after the expansion of molecular biology in the 1950's, it became evident that by comparing the proteins and nucleic acids of one species with those ...other, one could hope to obtain ...antitative and objective estimate ...e "genetic distance" between spe- Until then, there was no common

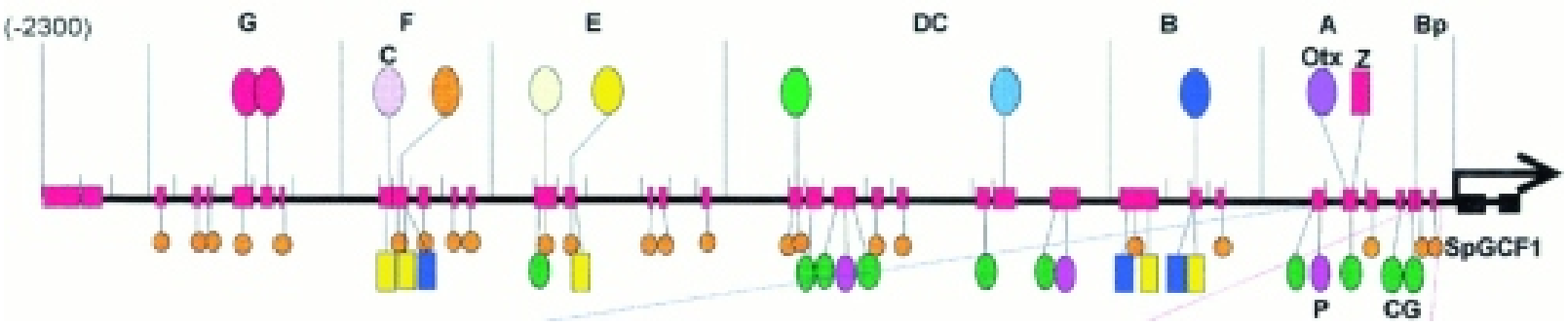


Summary and Conclusions

The comparison of human and chimpanzee macromolecules leads to several inferences:

1) Amino acid sequencing, immunological, and electrophoretic methods of protein comparison yield concordant estimates of genetic resemblance. These approaches all indicate that the average human polypeptide is more than 99 percent identical to its chimpanzee counterpart.

cis-regulatory sequences in eukaryotes



E. Davidson

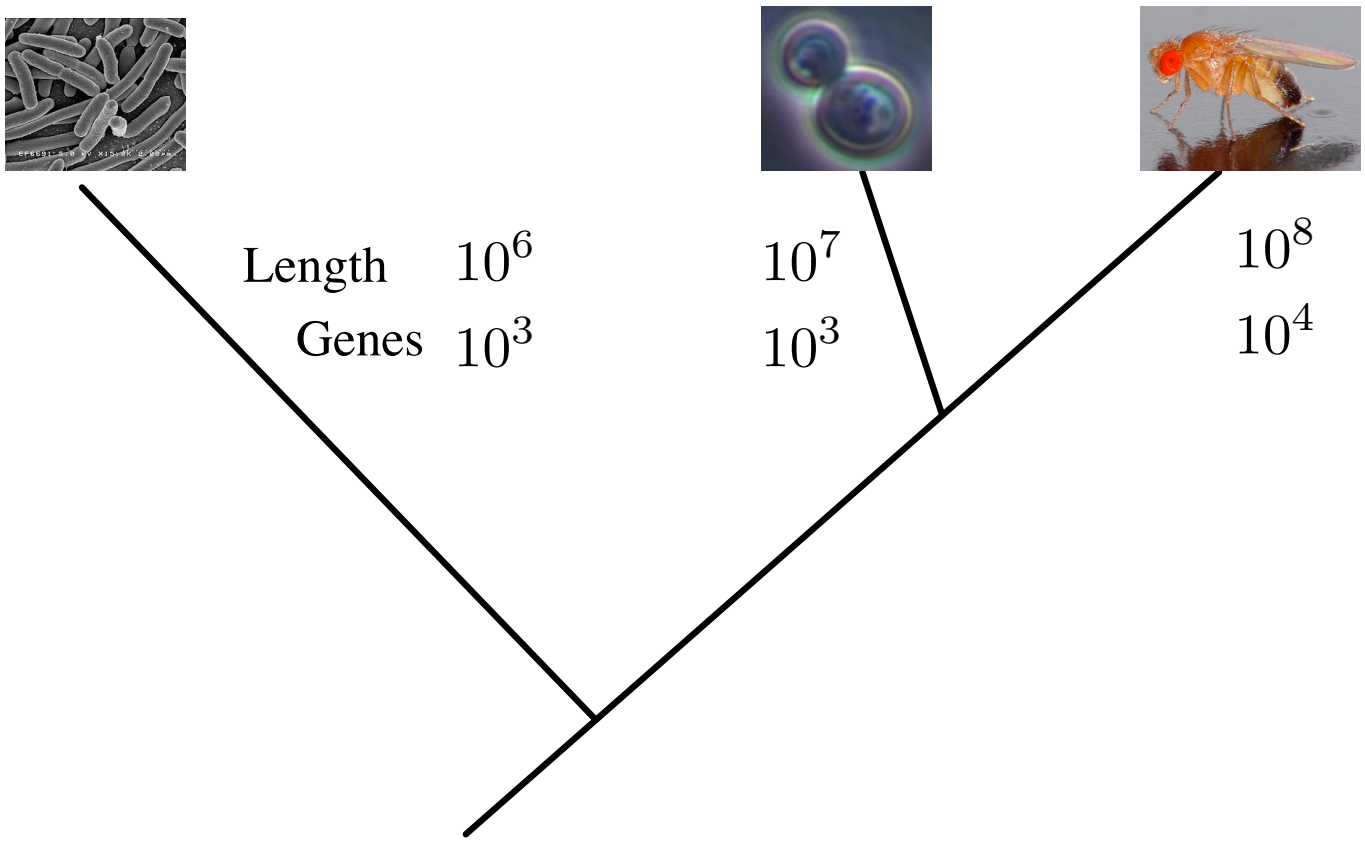
- Regulatory grammar and logic

  E. Davidson (See Urchin regulatory networks), D. Arnosti (billboard model, short range repression in enhancers), M. Levine, S. Small (Drosophila regulation)
- Functional conservation over large evolutionary distances

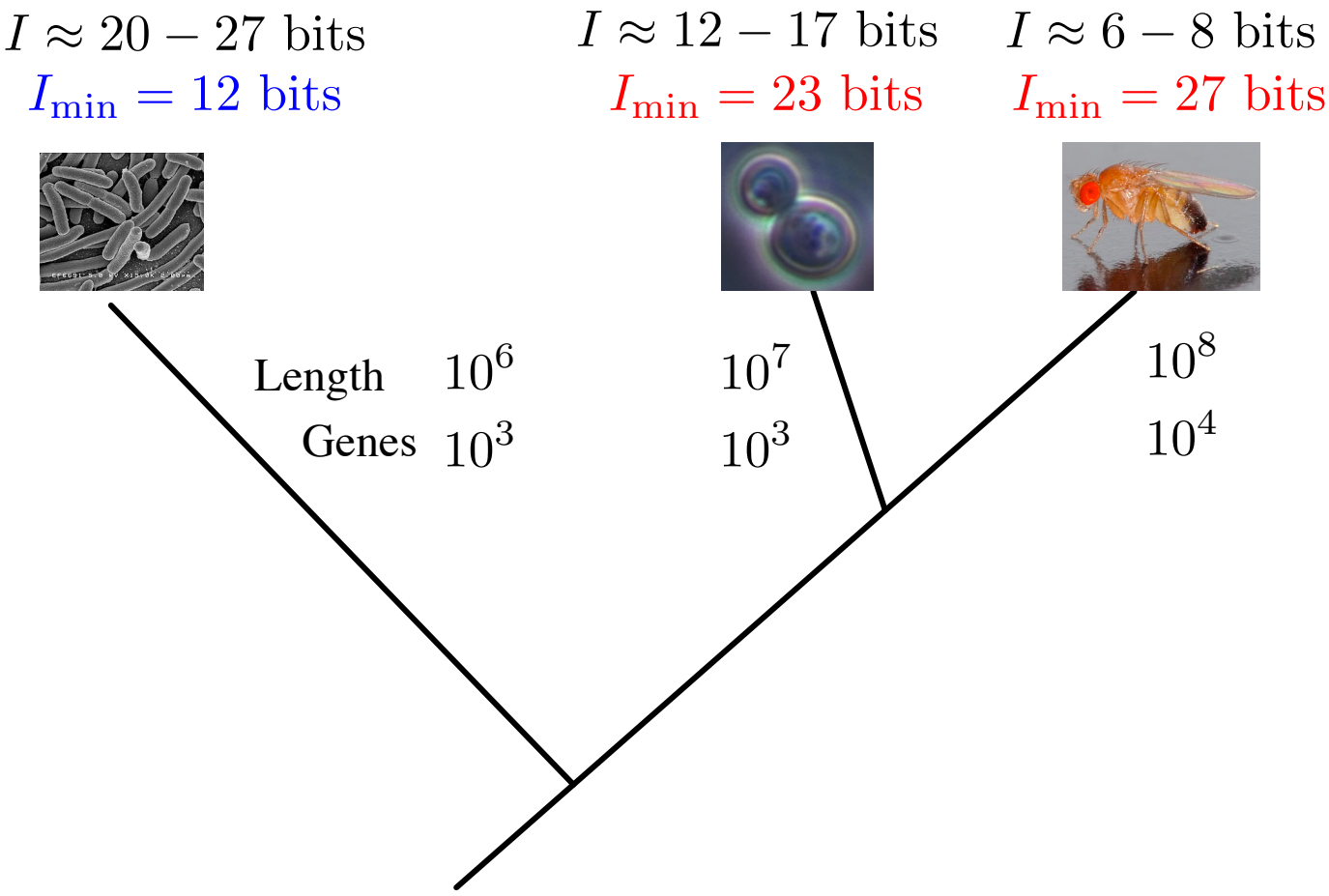
  Ludwig et al. (even-skipped in Drosophila), Hare et al (Sepsid-Drosophila comparison)
- High rates of sequence turnover inferred from comparative genomics, A. Moses et al. (2006), S. Doniger and J. Fay (2007)
- High quality databases on regulatory information

  REDfly, SwissRegulon and etc.

Encoding complexity



Encoding complexity



- Minimum information to identify a unique object among N alternatives

$$I_{\min} = \log_2 N$$

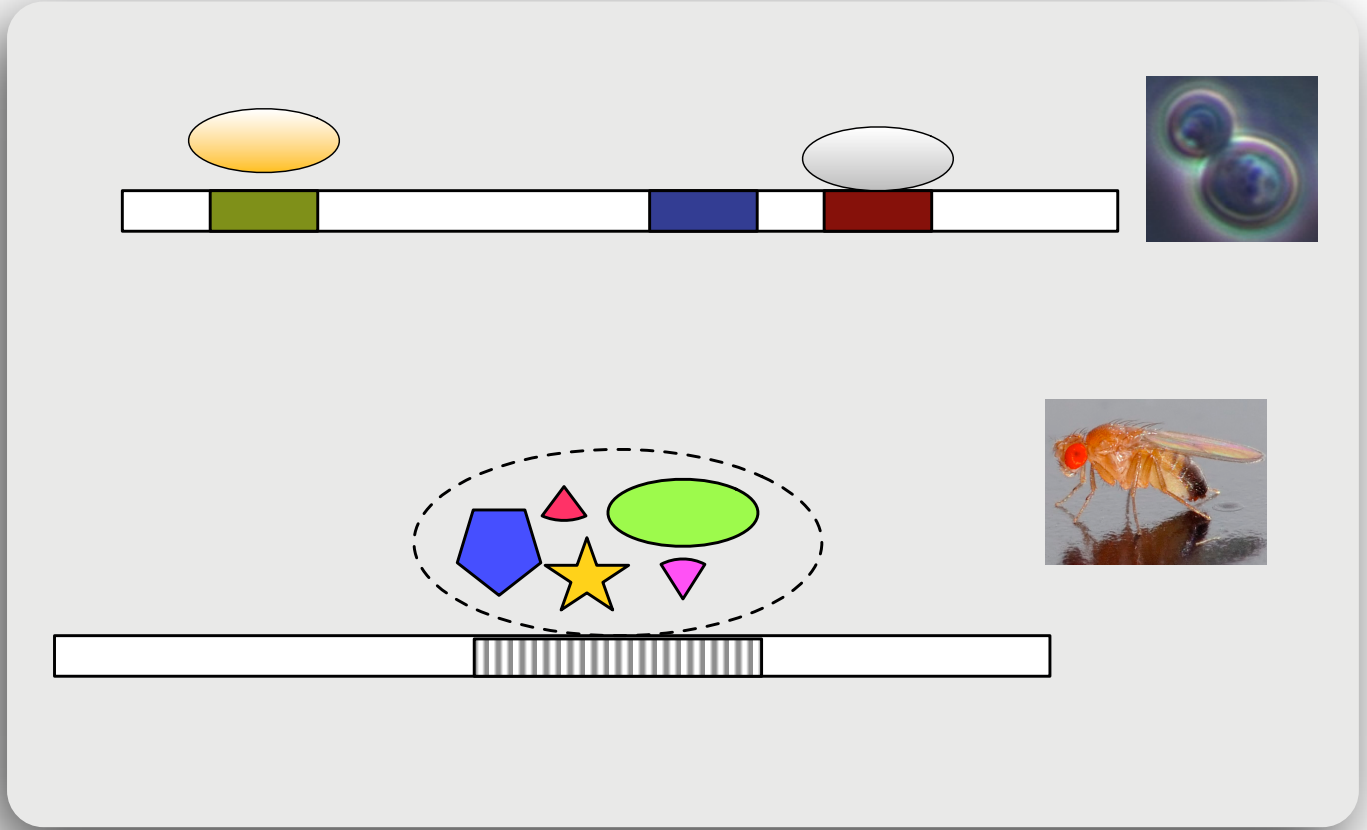
Z. Wunderlich, L. Mirny (2009)

Encoding complexity

$I \approx 20 - 27$ bits
$I_{\min} = 12$ bits

$I \approx 12 - 17$ bits
$I_{\min} = 23$ bits

$I \approx 6 - 8$ bits
$I_{\min} = 27$ bits



object among N alternatives

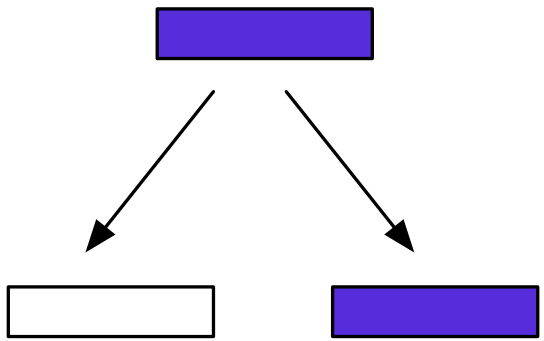
$I_{\min} = \log_2 N$

Z. Wunderlich, L. Mirny (2009)

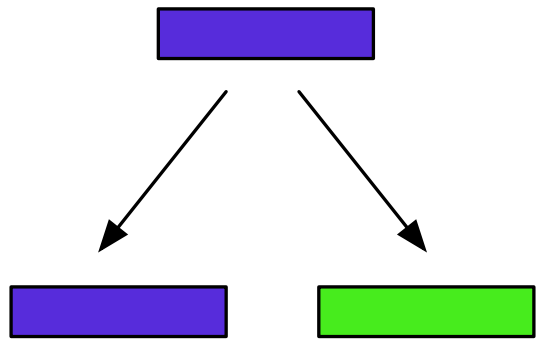
How does complexity evolve?

Functional diversification by gene duplication

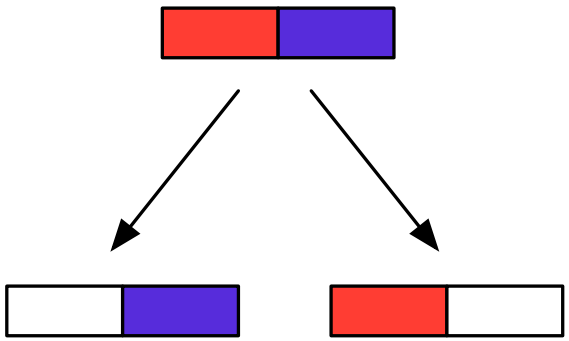
Non-functionalization



Neo-functionalization

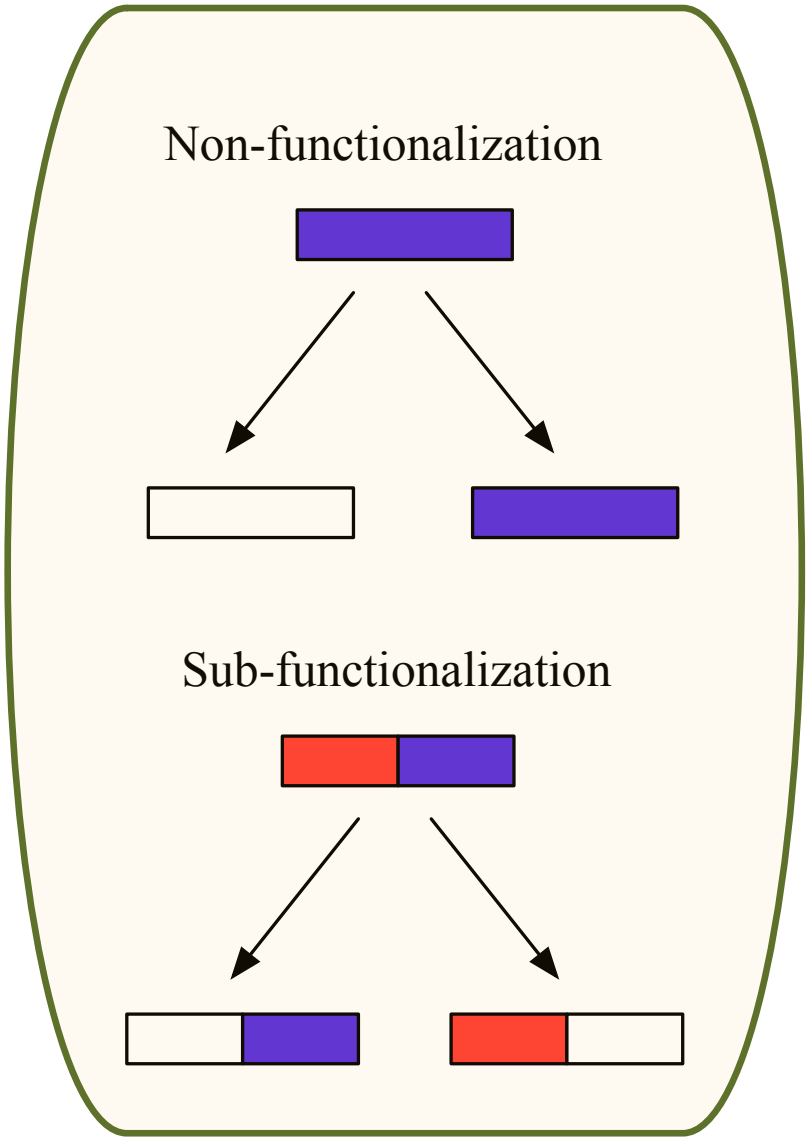


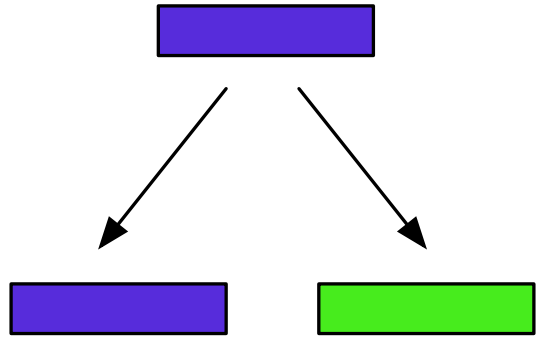
Sub-functionalization



M. Lynch, A. Force (2000)
M. Lynch, J. Conery (2003)

Functional diversification by gene duplication





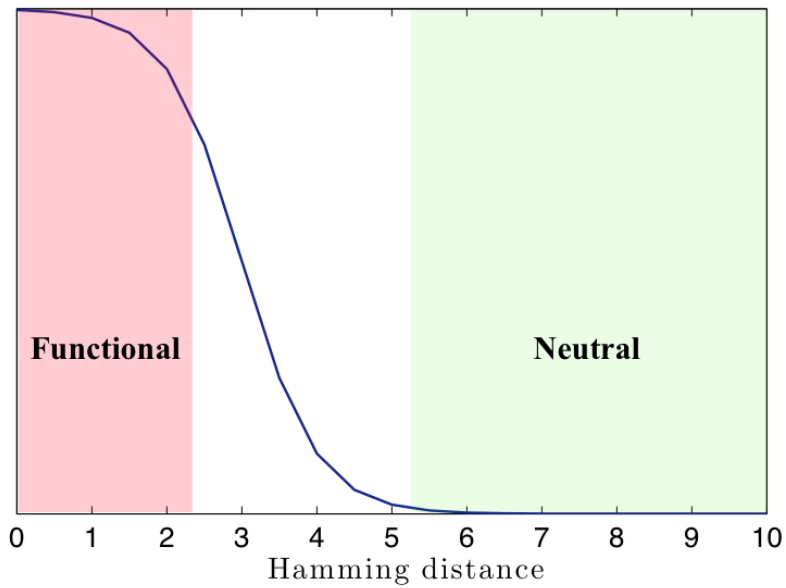
Loss of regulatory inputs per gene!
Reduction of promoter complexity!

M. Lynch, A. Force (2000)
M. Lynch, J. Conery (2003)

Formation of binding sites

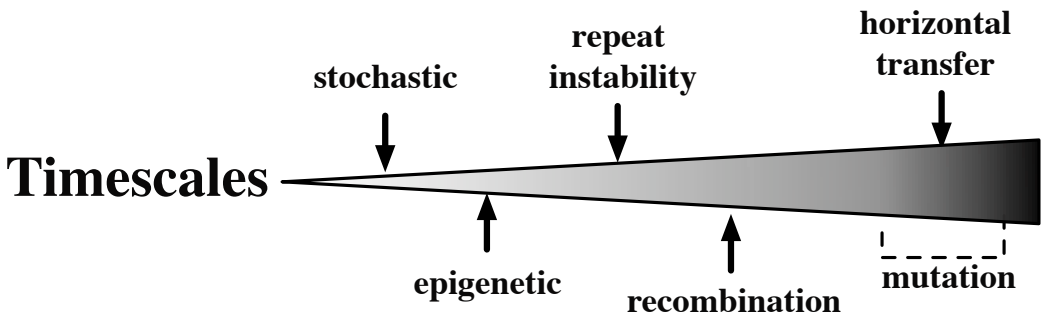
- Point mutations alone cannot explain the adaptive formation of regulatory clusters, *J. Berg et al. (2004)*

Biophysics of the interactions generates a cliff-type fitness landscape for factor binding



Short repeats in regulatory sequences

- Influence on regulatory function
- Transcriptional evolvability, Vinces et al (2009)
- Gaps in sequence alignments and short repeats
- Surplus of insertion events by short tandem repeats in the regulatory regions of *Drosophila*, S. Sinhe and E. Siggia (2005)



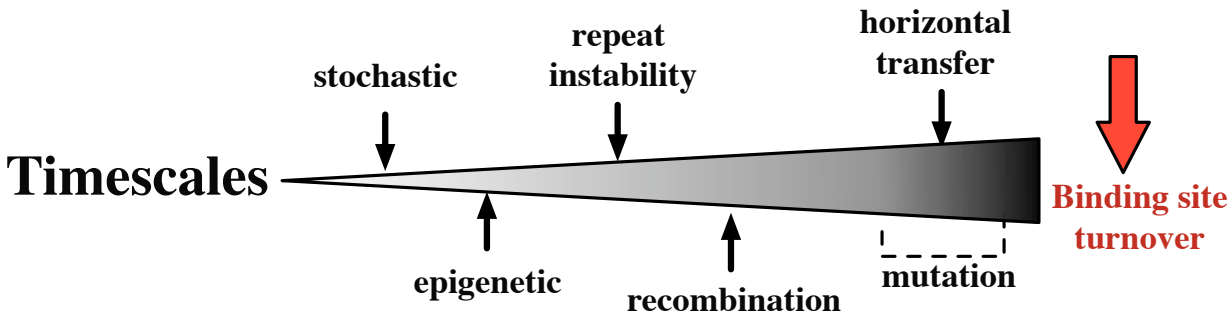
N. Maheshri

Short repeats in regulatory sequences

- Influence on regulatory function
- Transcriptional evolvability, Vinces et al (2009)
- Gaps in sequence alignments and short repeats
- Surplus of insertion events by short tandem repeats in the regulatory regions of *Drosophila*, S. Sinhe and E. Siggia (2005)

- Timescales for repeat evolution and binding site turnover are very different

Can indels confer regulatory information?



N. Maheshri

Which sequence evolution modes produce regulatory information?

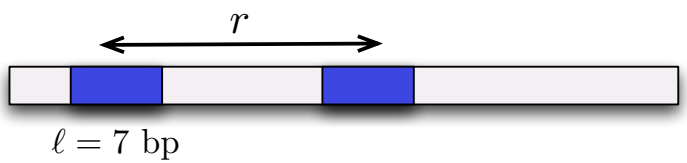
Traces of duplications in CRMs

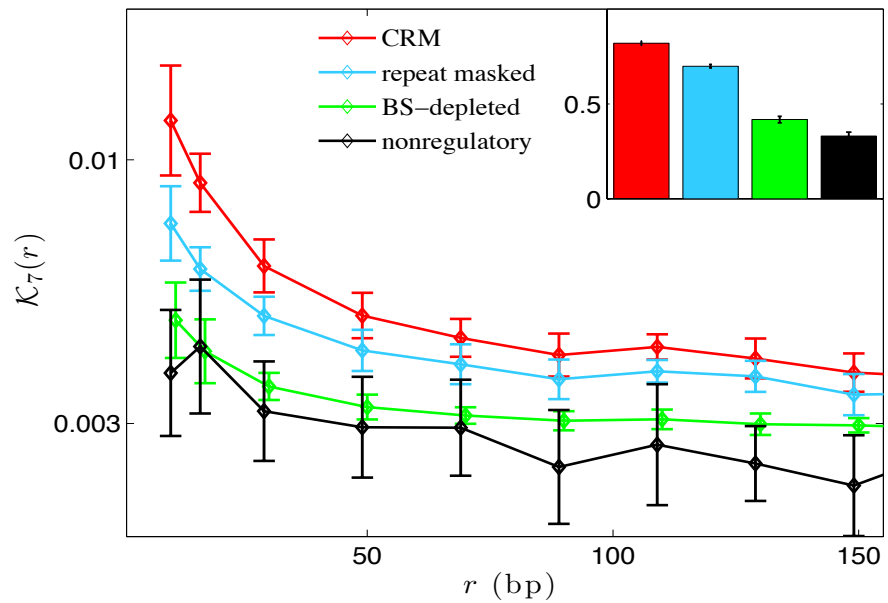
- Nucleotides in regulatory regions are correlated in the distance range of $r < 100$ bp.

Traces of duplications in CRMs

- Nucleotides in regulatory regions are correlated in the distance range of $r < 100$ bp.

- Mutually correlated nucleotides occur in local clusters with characteristic length of $\ell = 7$ bp.



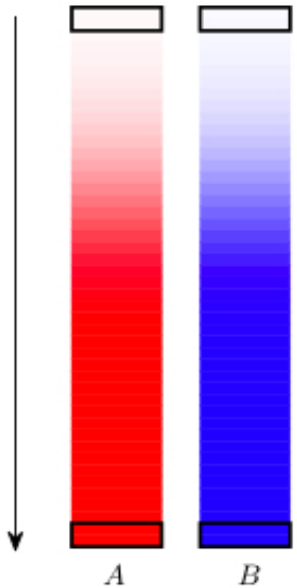


- Correlated binding sites explain a substantial part, microsatellite repeats only a small part of the similarity information.

A.N, M. Lässig (submitted)

Evolutionary modes of binding sites

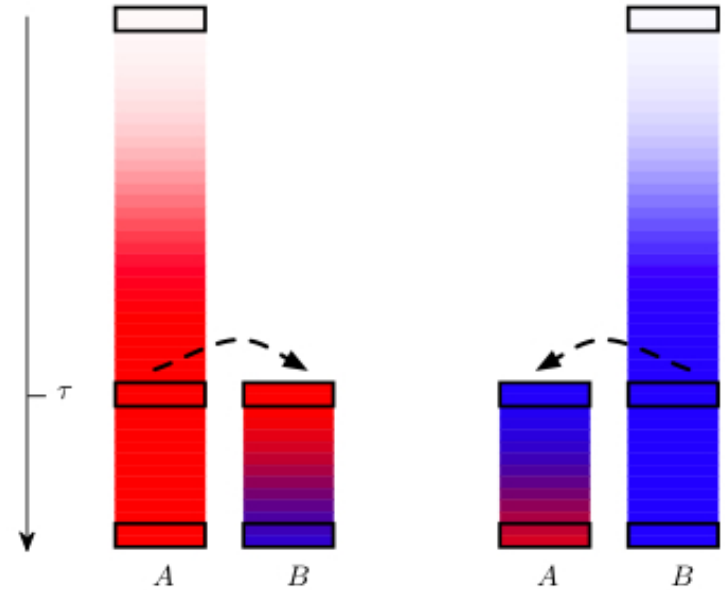
Independent Evolution



$Q^{\infty}(a,b) = Q_A(a)Q_B(b)$

NO enhanced sequence similarity compared to the motif ($S < 0$)

Evolution by Sequence Duplication



$Q^{\tau}(a,b) = \sum_c G_A^{\tau}(a|c)G_B^{\tau}(b|c)Q(c)$

Enhanced sequence similarity compared to the motif ($S > 0$)

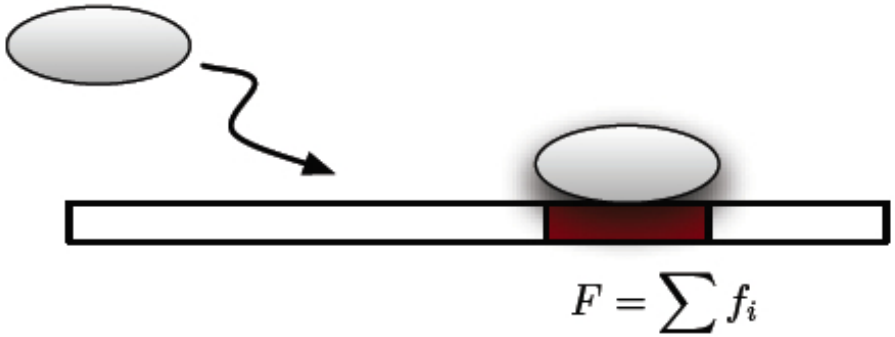
- We distinguish between the two evolutionary histories: $S^{\tau}(a,b) = \log \frac{Q^{\tau}(a,b)}{Q_A(a)Q_B(b)}$

A.N, M. Lässig (submitted)

Evolutionary model for binding sites

- Fitness landscape is derived from nucleotide frequencies of the sites (Halpern & Bruno, 1998)

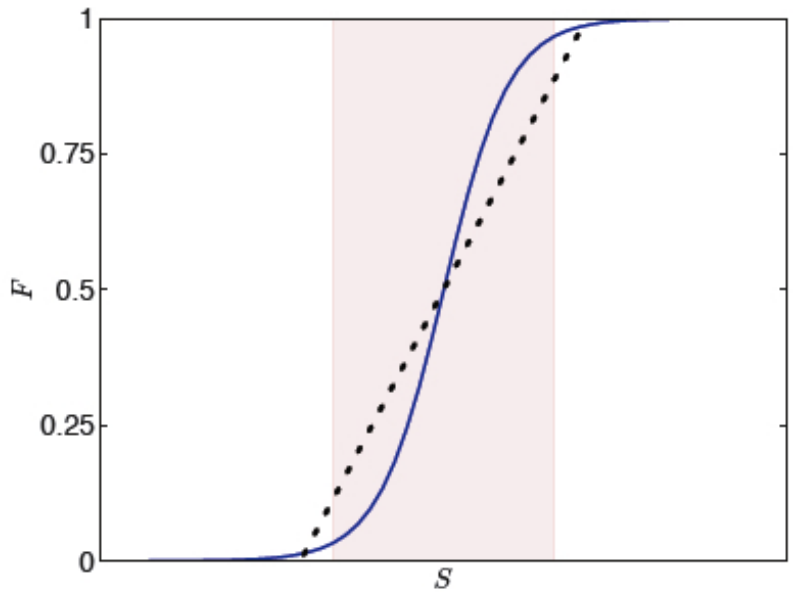
$$Q(a) = P_0(a)\mathrm{e}^{NF(a)}$$



- Mutation, selection and genetic drift drive the evolution of the binding sites

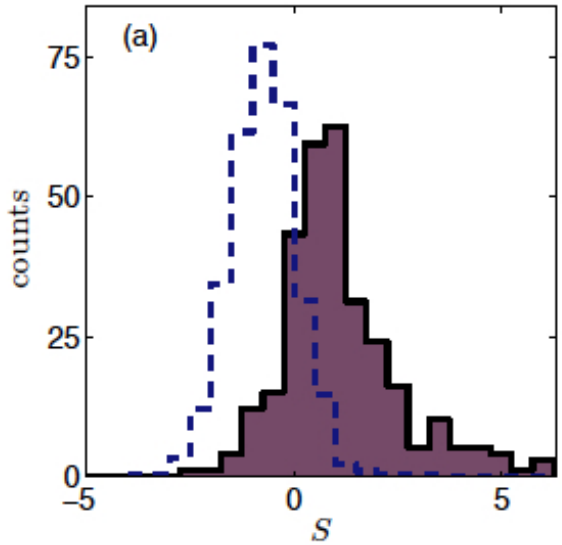
- Substitution rate (Kimura, 1967)

$$u_{a\to b} = \mu_{a\to b}\frac{N\Delta F_{ab}}{1-\exp(-N\Delta F_{ab})}$$



Binding site formation by local duplications

- Colseby binding sites share a common sequence ancestor in Drosophila.



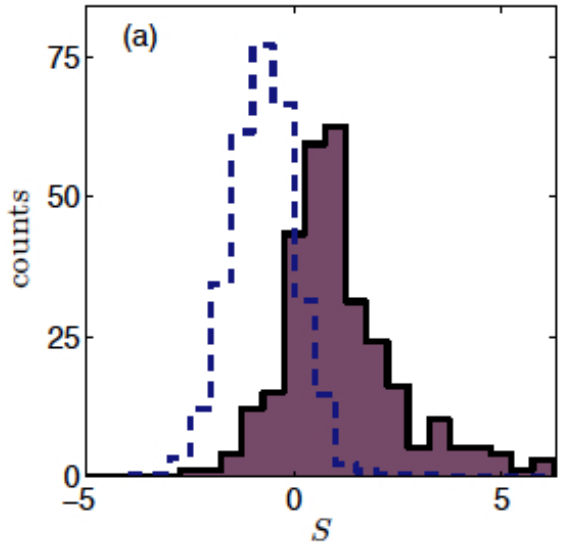
306 site pairs with $d < 50$ bp in D. mel

$\langle S \rangle = 1.3$, $\Sigma = 398$

Biased Dice

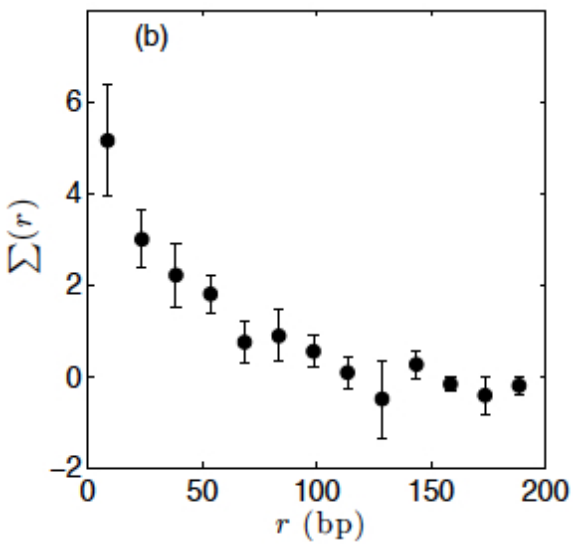
Binding site formation by local duplications

- Colseby binding sites share a common sequence ancestor in Drosophila.
- Sequence similarity is local.



306 site pairs with $d < 50$ bp in D. mel

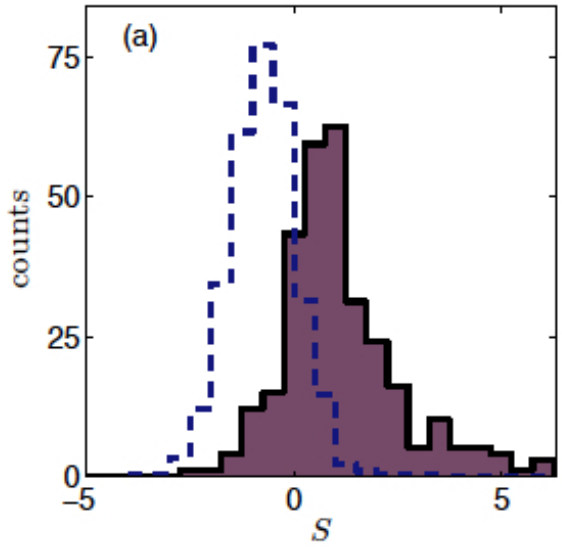
$\langle S \rangle = 1.3$, $\Sigma = 398$



Biased Dice

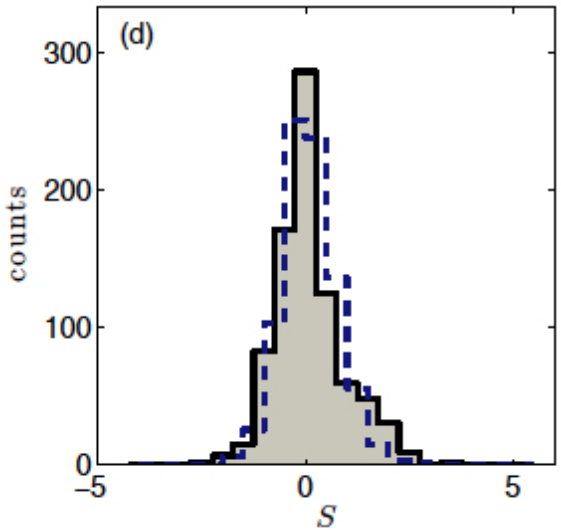
Binding site formation by local duplications

- Colseby binding sites share a common sequence ancestor in Drosophila.
- Sequence similarity is local.
- Common descent is not the prevalent evolutionary mode in yeast.

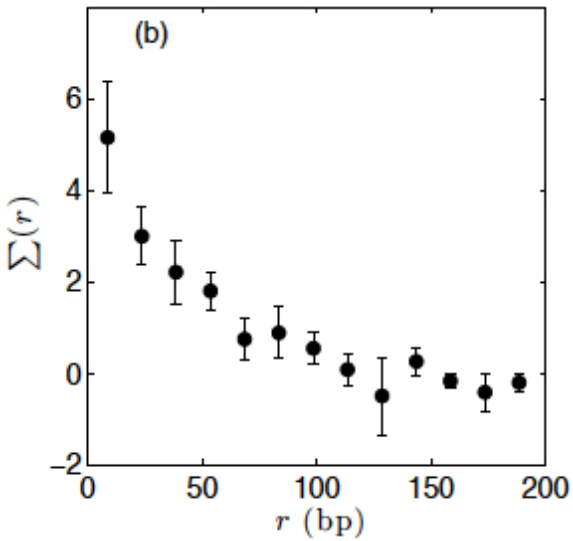


306 site pairs with $d < 50$ bp in D. mel

$\langle S \rangle = 1.3$, $\Sigma = 398$



833 site pairs with $d < 50$ bp in S. cere



Discussion

- Asymmetric life cycle of binding sites in regulatory modules
  - **Formation** by local duplication in clusters
  - Adaptation by point mutation
  - Optimization of the relative distance by indels
  - Conservation by stabilizing selection
  - **Loss** by point mutations

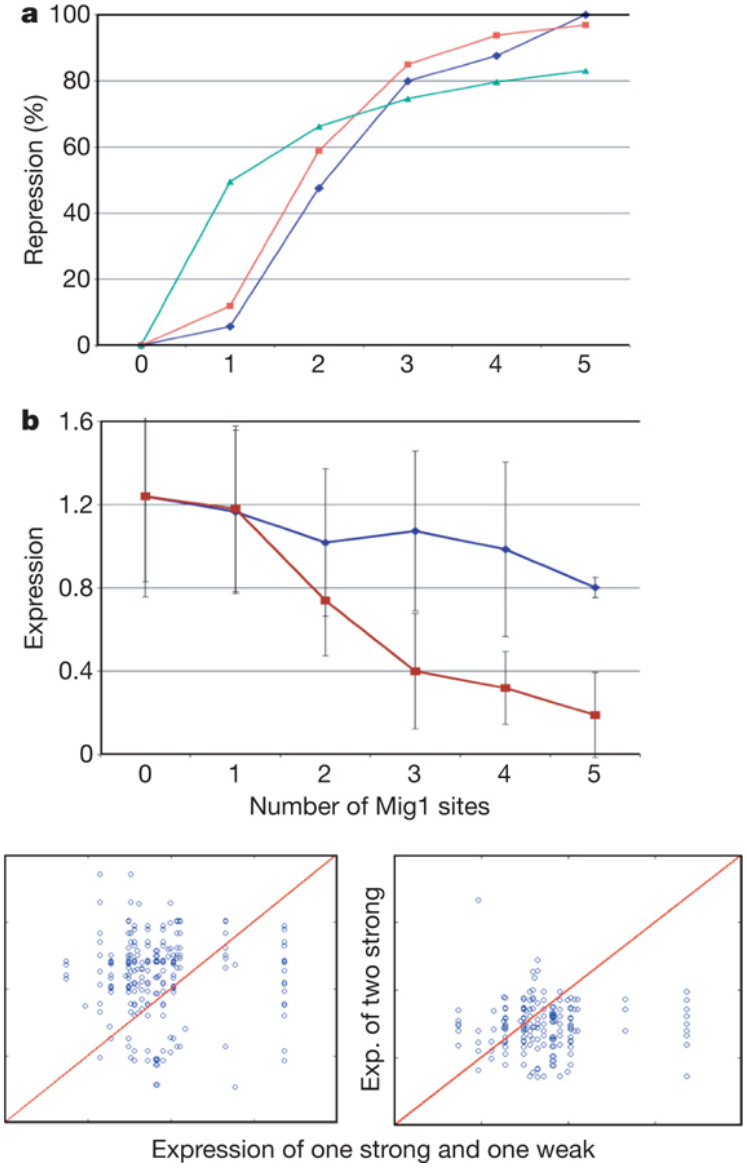
- Modes of sequence evolution and regulatory grammar

  What is the result and what is the substrate?

Shadow of weak binding sites

The role of weak binding sites in regulation

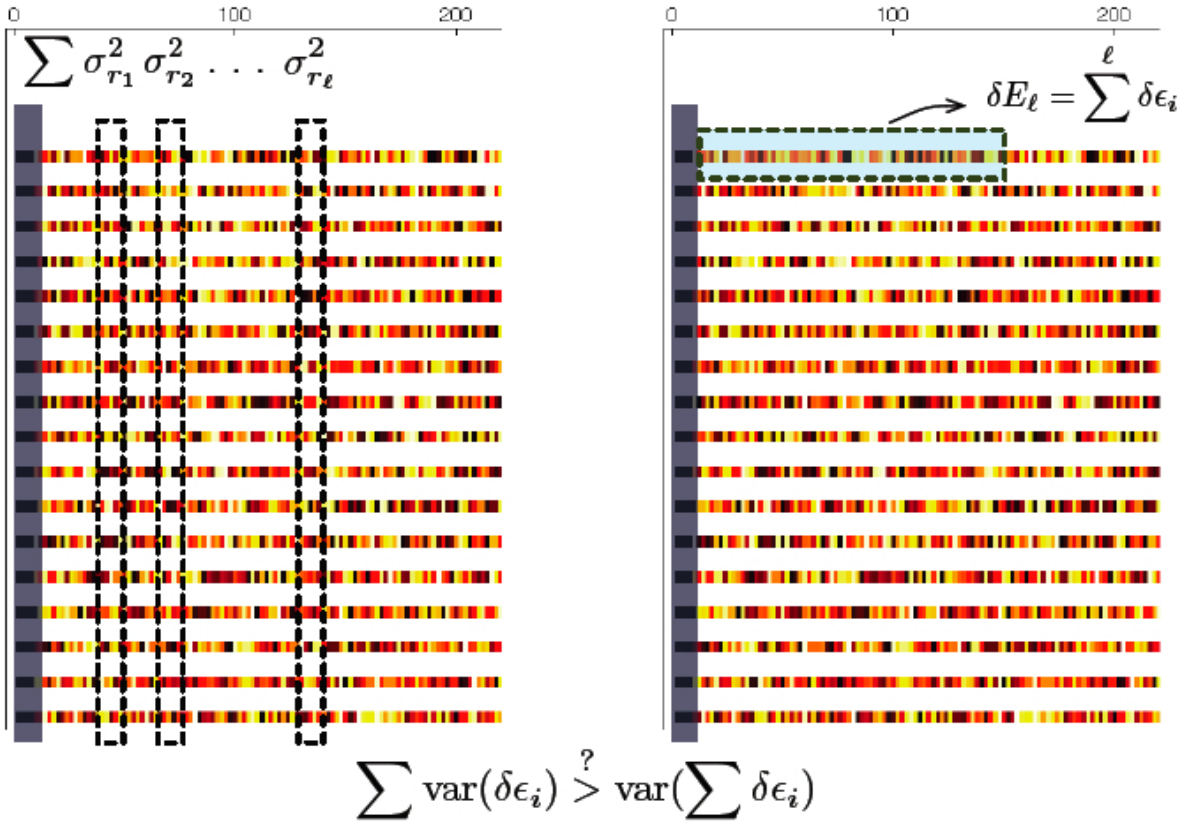
- Mig1-binding sites act cooperatively (Hill coefficient 3.4)
- Weak Mig1 sites repress weakly
- One weak site can be sufficient in cooperation with a strong site



Gertz et. al., Nature 2009

Stabilizing selection in yeast

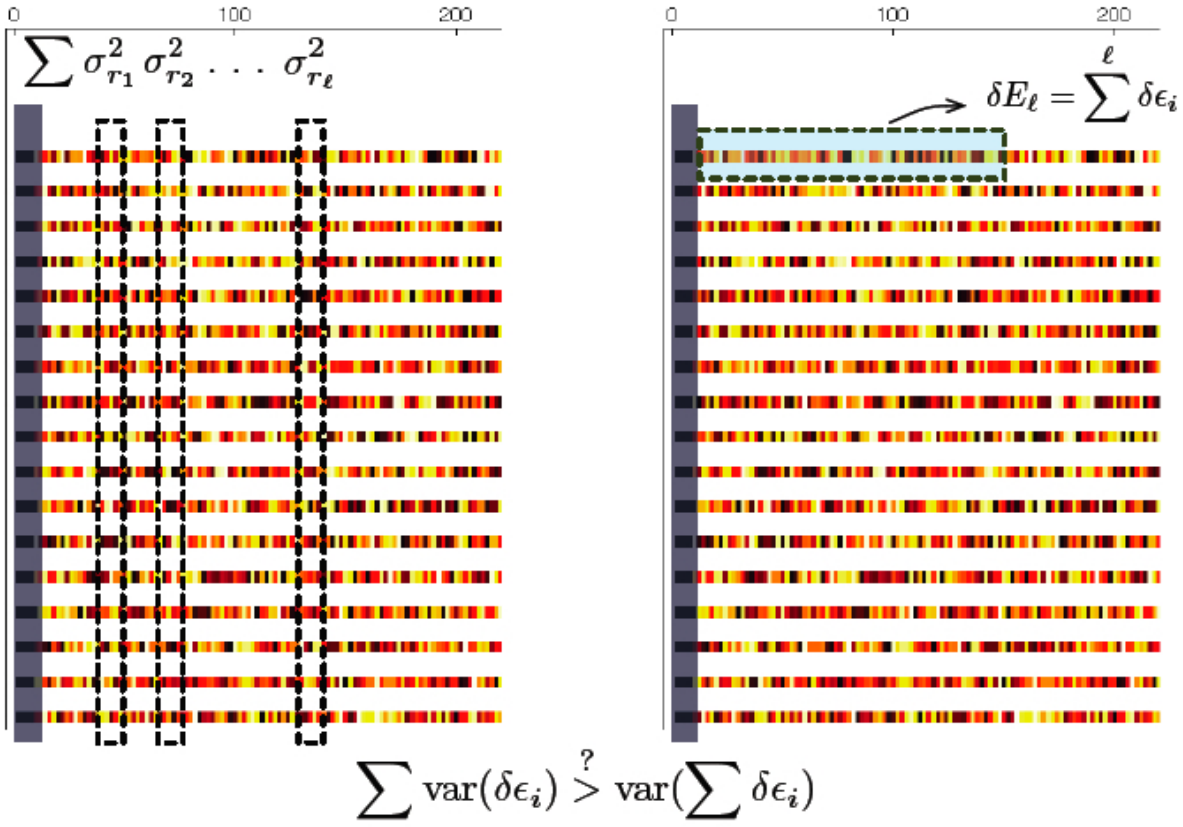
Can we see compensatory evolution?

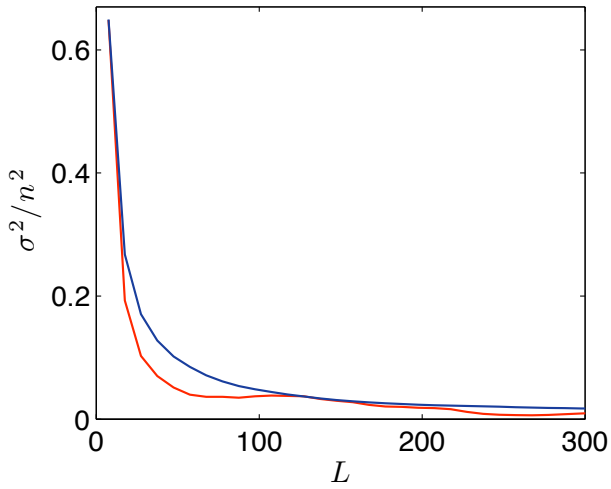


$$\sum \mathrm{var}(\delta\epsilon_i) \overset{?}{>} \mathrm{var}(\sum \delta\epsilon_i)$$

Stabilizing selection in yeast

Can we see compensatory evolution?





Rap1 factor in yeast

MITOMI energy matrix (Maerkl lab)

- We see evidence for compensatory evolution of weak sites in the vicinity of a strong binding site.

Conclusion & Outlook

- Binding site clusters are mainly formed by local sequence duplications
- Local duplications can explain the asymmetric life-cylce of the binding sites
- Binding site duplications have adaptive advantage
- This type of duplications is not the prominent mode of site formation in yeast

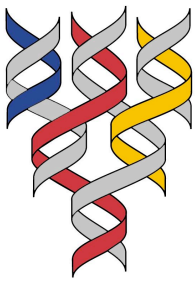
- Characterizing the promoter sequences as single entities (transition from single particle to many-particle statistics)
- Evolution of the resulted quantitative trait (epistatic, linked genome)

Thanks

Michael Lässig

Ville Mustonen (Sanger institute, Cambridge)

Sebastian Maerkl (EPFL)



SFB 680
Molecular Basis of
Evolutionary Innovations

bcgs

Bonn-Cologne Graduate School
of Physics and Astronomy