Molecular Information Theory of Composite Sequence Motifs

Elia Mascolo[1], Ivan Erill[1,2]

[1] University of Maryland Baltimore County, Baltimore, MD 21250, USA
[2] Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

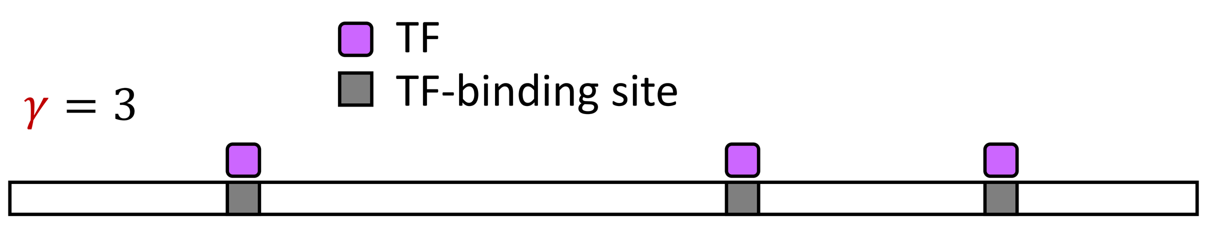
ErillLab UMBC

Classical Theory for Sequence Motifs

Information as a decrease in *uncertainty* (entropy): $H_{before} - H_{after}$

number of positions in the genome — *number of target sites*

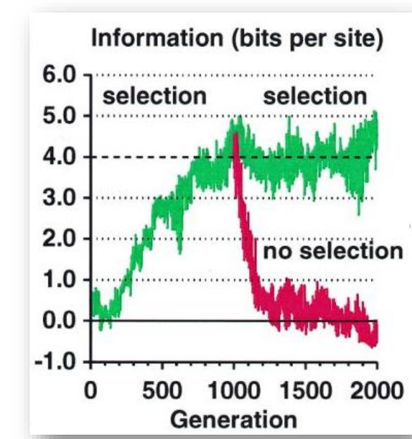
$$R_{frequency} = log_2(G) - log_2(\gamma) = -log_2\left(\frac{\gamma}{G}\right) \quad (bits)$$



$R_{frequency}$: information required to specify γ target sites on a genome of G bp.

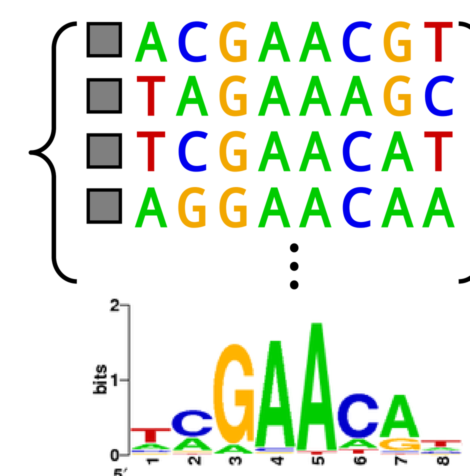
$R_{frequency} \approx R_{sequence}$

(Schneider *et al.*, 1986)



(Schneider, 2000)

The information the transcription factor needs is encoded in the DNA sequence of its binding sites



Aligned TF-binding sites

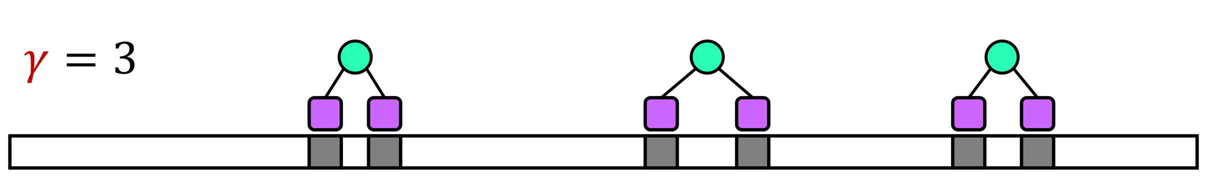
Information (*bits*) in each position of the motif

$R_{sequence}$: information contained in the motif

Composite Motifs ($n = 2$)

number of dimer placements — *number of target placements*

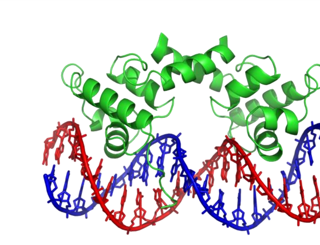
$$R_{frequency} = log_2(G^2) - log_2(\gamma) = -log_2\left(\frac{\gamma}{G^2}\right) \quad (bits)$$



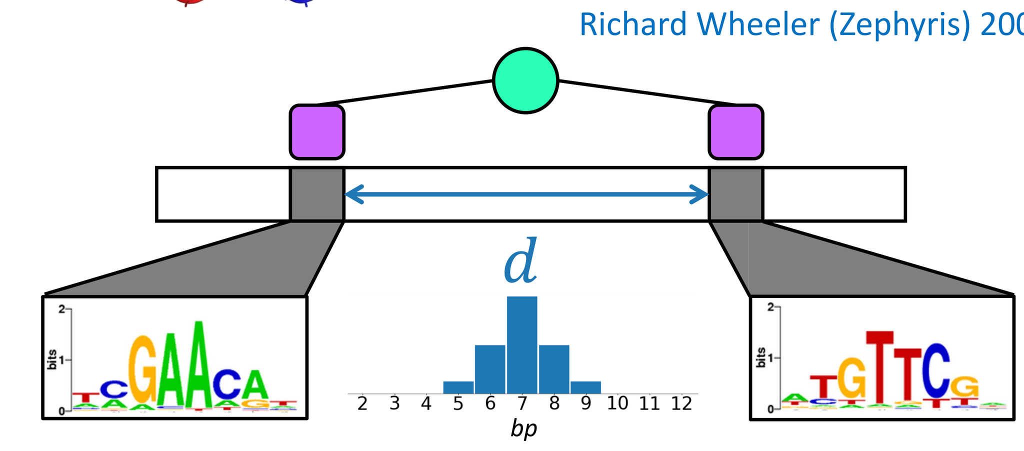
$$-log_2\left(\frac{\gamma}{G^2}\right) \approx$$
$$\approx R_{sequence(1)} + R_{sequence(2)} + R_{spacer}$$

$$R_{sequence(1)} \leq -log_2\left(\frac{\gamma}{G}\right)$$
$$R_{sequence(2)} \leq -log_2\left(\frac{\gamma}{G}\right)$$
$$log_2(\gamma) \leq R_{spacer} \leq log_2(G)$$

A TF dimer ($n = 2$)



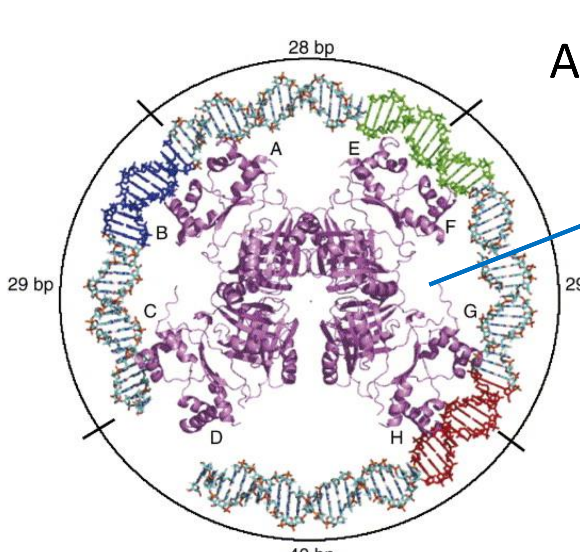
Lambda repressor (Wikimedia Commons - Author: Richard Wheeler (Zephyris) 2007)



$R_{sequence(1)}$ R_{spacer} $R_{sequence(2)}$

General Theory ($n \geq 1$)

A TF octamer ($n = 8$)



Leucine-responsive regulatory protein Lrp (de los Rios and Perona, 2007)

A GENERAL INFORMATION THEORY OF COMPOSITE MOTIFS

(equivalent to Schneider's equation in the special case when $n = 1$)

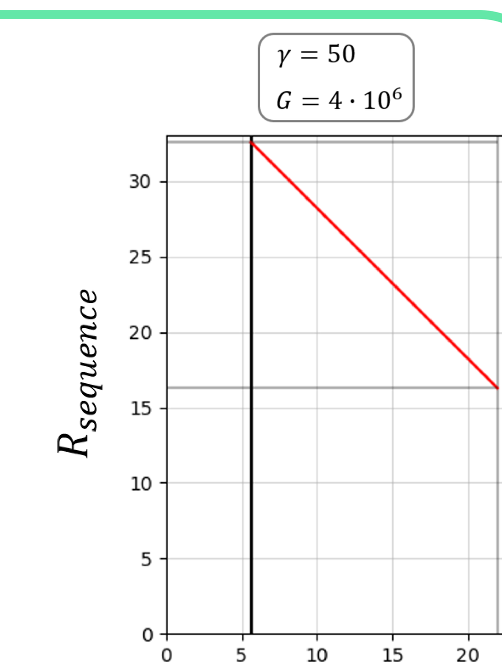
$$R_{frequency} = log_2(G^n) - log_2(\gamma) = -log_2\left(\frac{\gamma}{G^n}\right) \quad (bits)$$

$$\sum_{i=1}^{n} R_{sequence(i)} + \sum_{i=1}^{n-1} R_{spacer(i)} \approx -log_2\left(\frac{\gamma}{G^n}\right)$$

$$R_{sequence(i)} \leq -log_2\left(\frac{\gamma}{G}\right) \ \forall\, i \leq n$$
$$log_2(\gamma) \leq R_{spacer(i)} \leq log_2(G) \ \forall\, i \leq n-1$$

We can re-write it by redefining the terms as:

$$R_{sequence} + R_{spacer} \approx R_{frequency}$$



Regulator's biophysics

Harmonic oscillator in thermal bath → The distance between recognizers is Gaussian, with variance:

$$\sigma_{protein}{}^2 = \frac{k_B T}{\kappa}$$

spring constant (stifness)

κ_{opt}: value of κ such that

$$\sigma_{protein}{}^2 = \sigma_{targets}{}^2$$

Energy dissipation

Landauer's limit:

$$E_{min} = k_b T\, ln(2) \quad (joules\ per\ bit)$$

Minimum energy dissipation per target recognition:

RECRUITMENT

$k_b T ln(2)(R_{sequence} + R_{spacer})$ *joules*

Recruitment-based searches: less thermodynamically efficient, but ≥ 2 possible output states (combinatorial control).

PRE-RECRUITMENT

$k_b T ln(2)(R_{sequence})$ *joules*

when the flexibility of the protein structure matches the spacer size distribution (e.g., Gaussian spacers: $\sigma_{protein}{}^2 = \sigma_{targets}{}^2$)

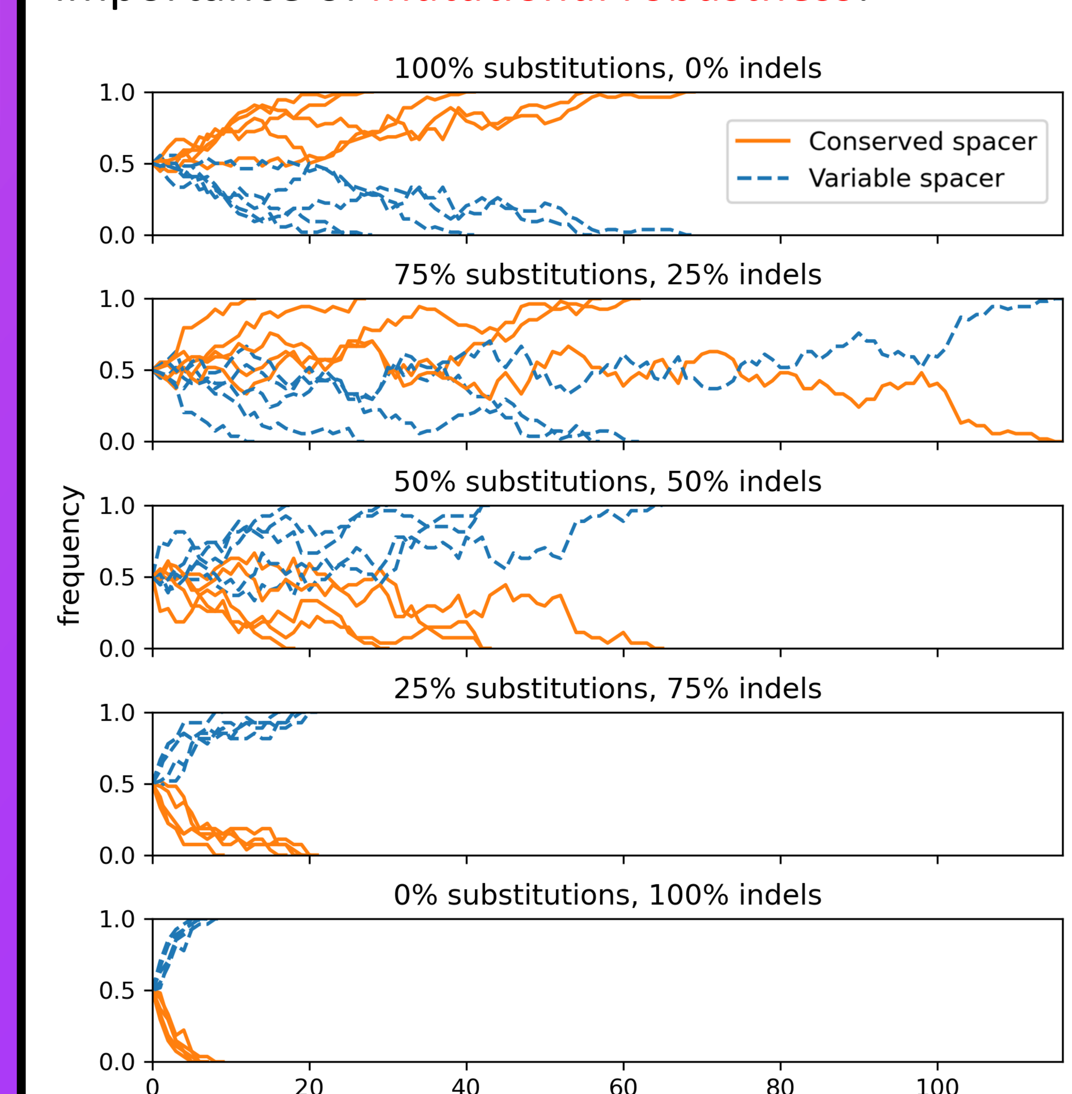
Competing strategies

$R_{sequence}$ VS R_{spacer}

What encoding strategy should be prioritized? It depends on mutation rates:

substitutions VS *indels*

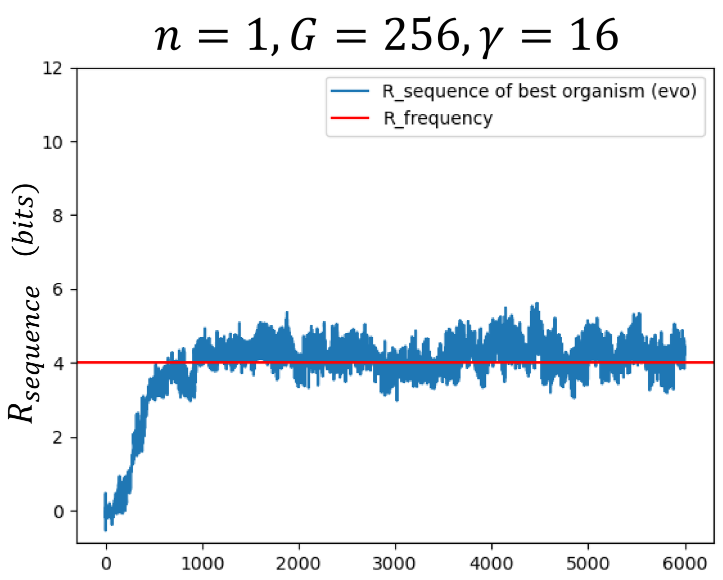
Competition experiments demonstrate the importance of *mutational robustness*.



Evolutionary simulations

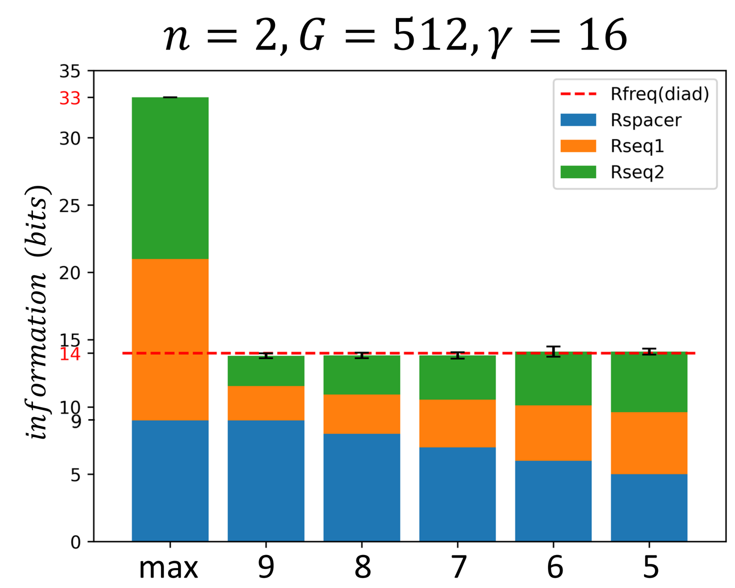
$n = 1$ to reproduce results from (Schneider, 2000)

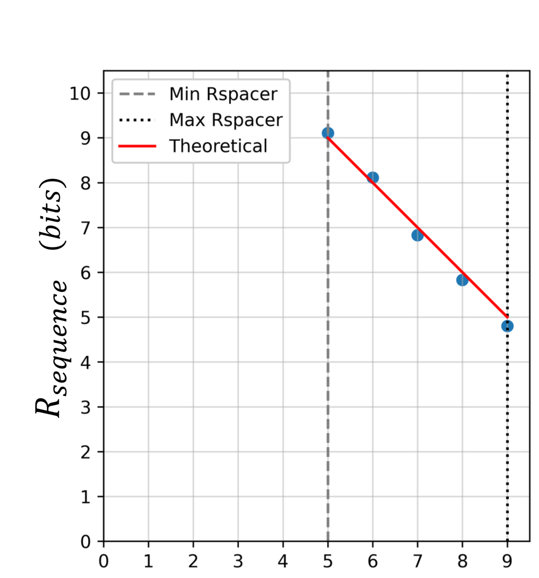
$n = 1, G = 256, \gamma = 16$



$n = 2$ to validate the relationship between $R_{sequence}$ and R_{spacer} in composite motifs.

$n = 2, G = 512, \gamma = 16$





The proteins quickly evolve their flexibility to match the variability in the targets' spacer.