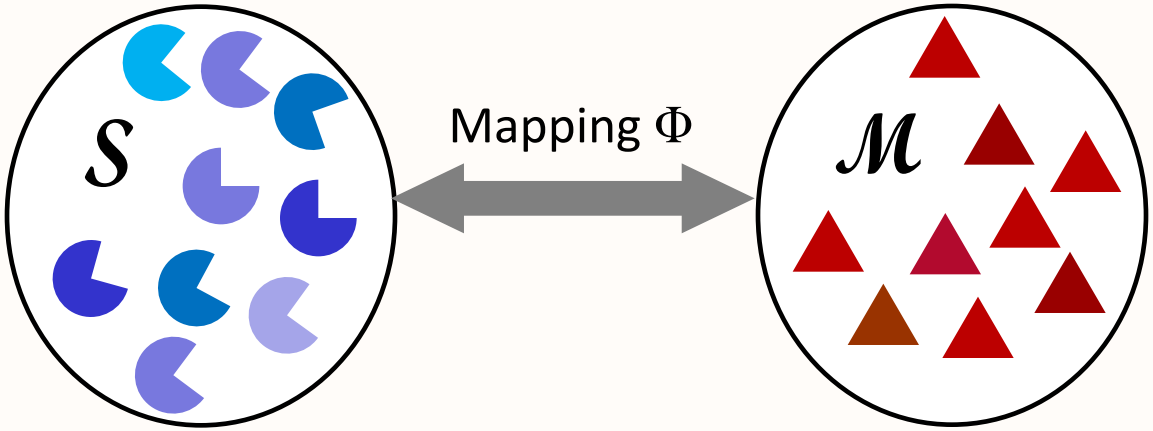
Physical Principles of Molecular Information Systems

Physics Colloquium

Weizmann

(G. Brodsky Mol Cell 2010)

Noisy molecular channels: Rate-Distortion theory

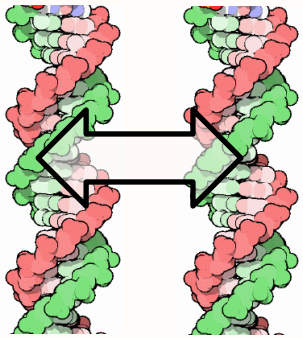


- Molecular spaces: $\mathcal{S}$, $\mathcal{M}$.
- Mapping: Φ: $\mathcal{S} \leftrightarrow \mathcal{M}$.
- Functionals (map): $\mathcal{F}(\Phi) \rightarrow$ "**fitness**"
- Optimize(fitness) : $\Phi^* = \text{argmax}\ \mathcal{F}(\Phi)$

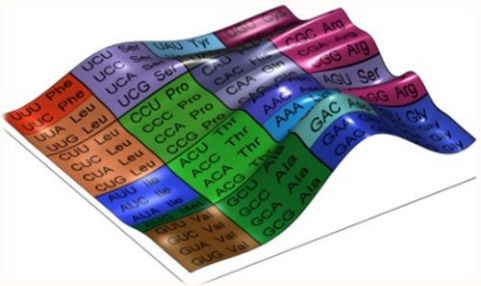
  $\rightarrow$ Design principles, Coding transition…
- Topological aspects.

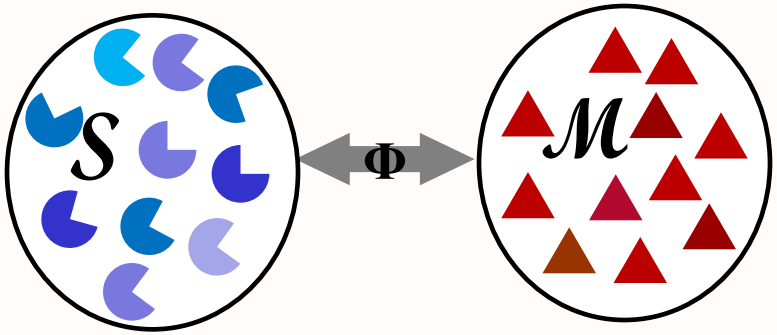
Examples: four fundamental living systems

Homologous Recombination

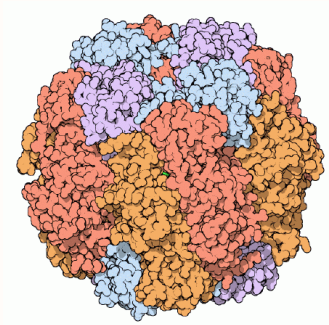


Molecular codes (genetic code)

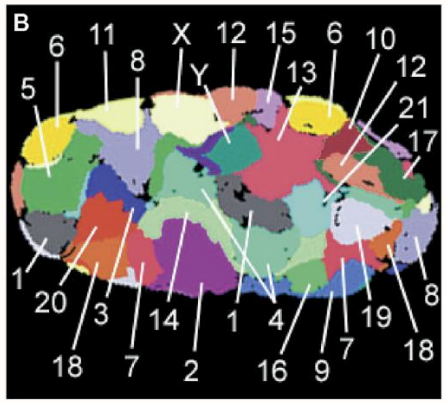




Photosynthesis

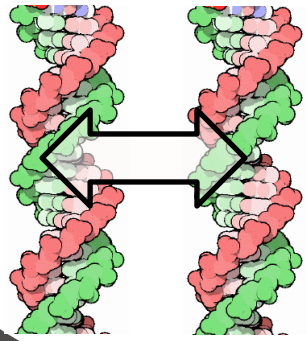


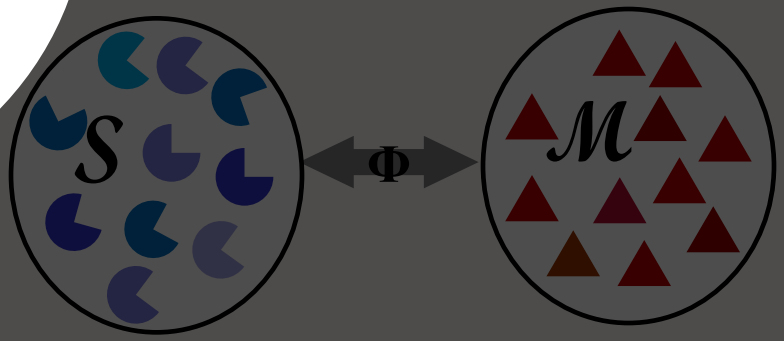
Chromosome organization



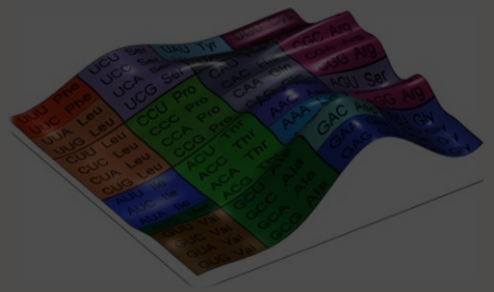
Examples: four essential living systems

Homologous Recombination

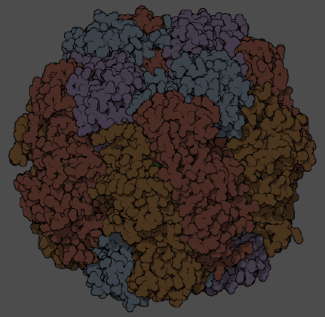




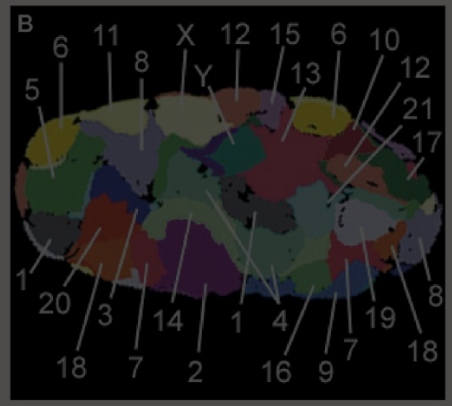
Molecular codes (genetic code)



Photosynthesis



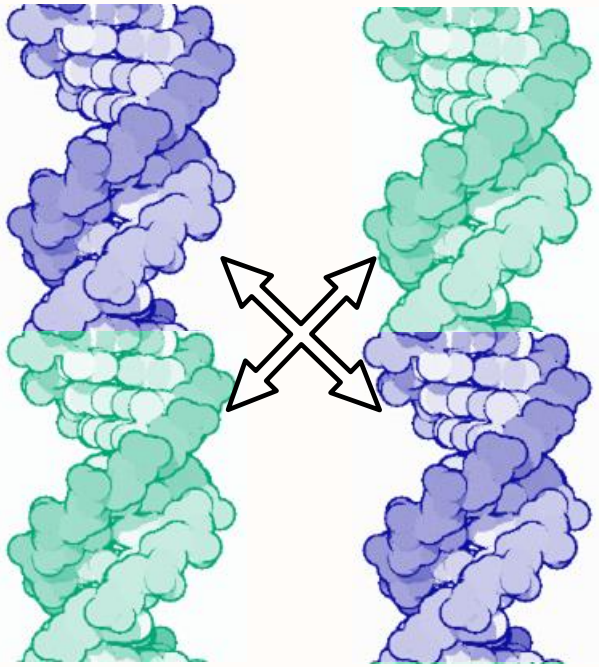
Chromosome organization



Recombination machinery recognizes homologous DNA

- Exchange between two homologous DNAs.
- Essential for:
  - *Genome integrity* (repair machinery).
  - *Genetic diversity* via crossover and sex (horizontal transfer).
  - affects speciation.

- Task: Detect correct, homologous DNA target among many incorrect lookalikes.



D. Goodsell

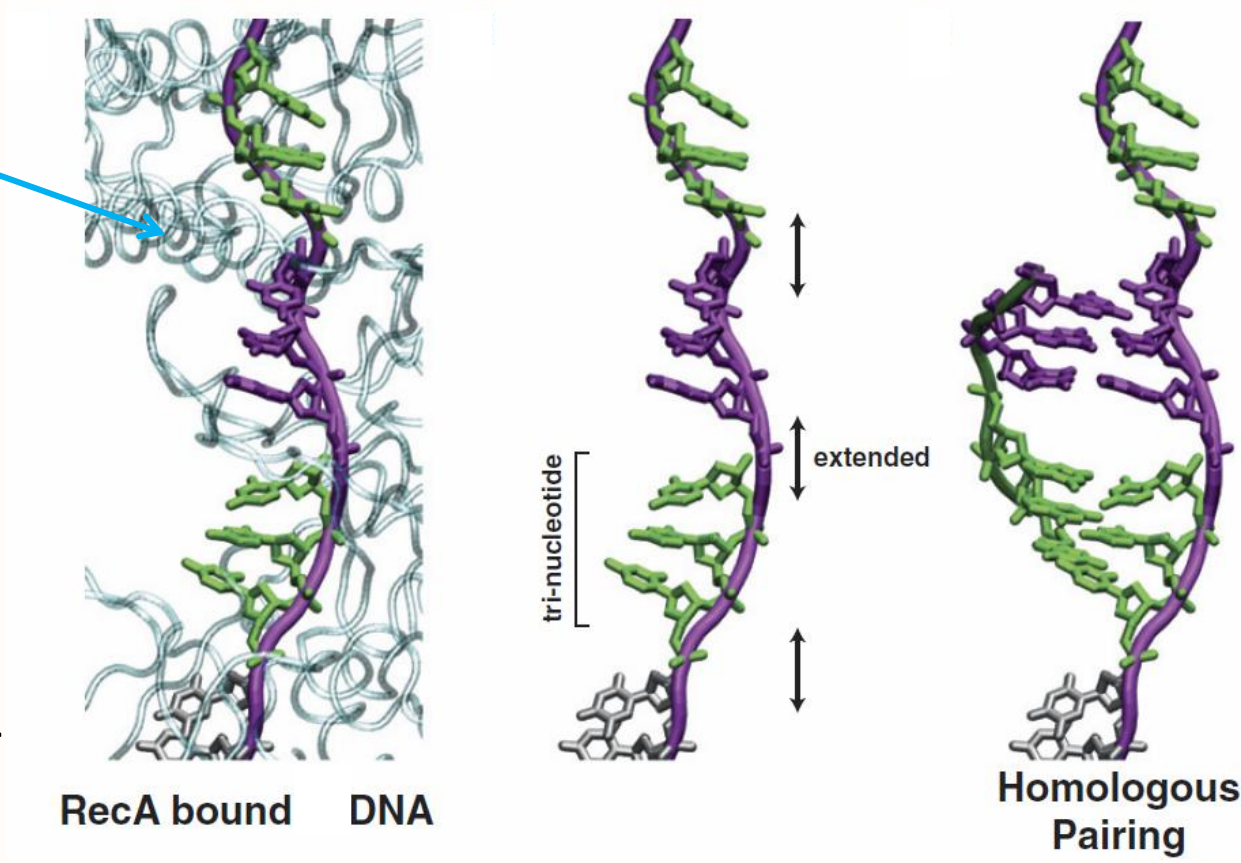
(Savir &TT, Plos 1 2007, Mol Cell 2010)

Why DNA is extended in homologous recombination?

- Mediated by **RecA.**
- Extension by 50%.
- Costs 3-4 k_BT/bp.

?

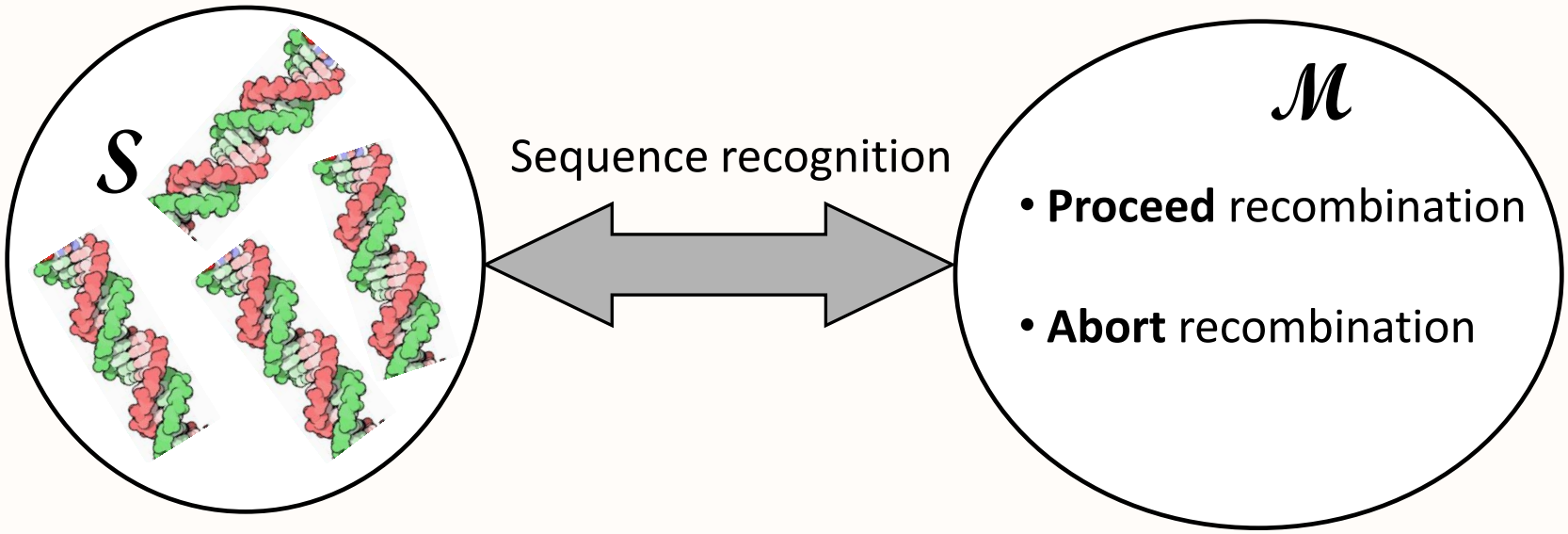
- mechanism to detect Right DNA in large pool of similar targets.
- General mechanism: Conformational proofreading.



(Savir & TT, Mol Cell 2010)

Homologous recombination maps sequences to decisions

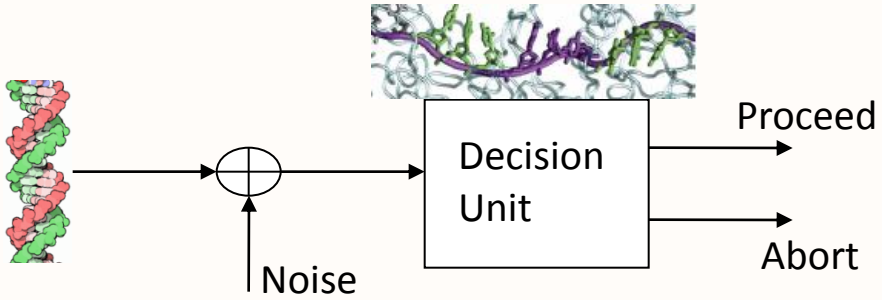
- Recombination advances in 3-base steps.
- Each step Mapping from DNA target space to decisions:



- Problem: Optimization to withstand noise in sequence recognition.

Molecular recognition is a decision problem

- Each 3-base step is **decision-making process**: 2x2 channel.



Output \ Input	Complementary Sequence (+)	Non-complementary Sequence (–)
Proceed with HR (+)	True Positive (+,+) $p_{+,+}$; $C_{+,+}$	False Positive (+,–) $p_{+,-}$; $C_{+,-}$
Abort HR (–)	False Negative (–,+) $p_{-,+}$; $C_{-,+}$	True Negative (–, –) $p_{-,-}$; $C_{-,-}$

- Each event = (input, output) bears cost/benefit, C(event) .
- **Fitness:** $\mathcal{F} = \sum_{\text{event}} C(\text{event}) \times P(\text{event}).$
- Fitness depends on structural parameters and can be optimized.

(Savir &TT, Plos 1 2007, Mol Cell 2010)

Extension maximizes recognition fitness

- Fitness depends on energetics via binding probabilities :

$$\mathcal{F}(C, P_b(E_b)) \approx P_{comp} - P_{non-comp}$$

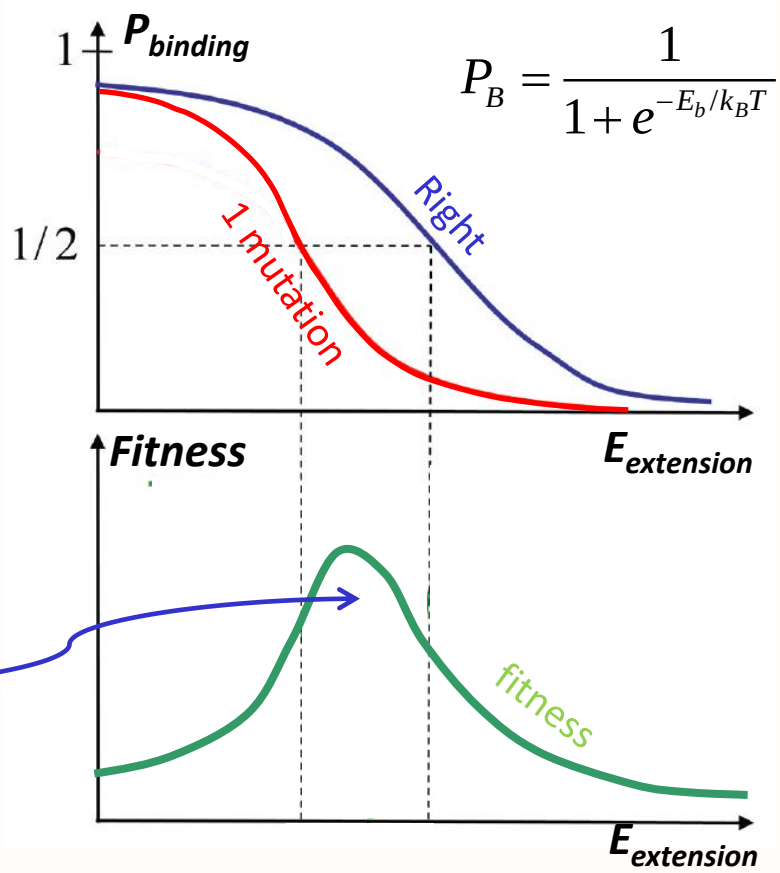
- One effective d.o.f. = $E_{\text{extension}}$

$$E_b = E_{pairing} - E_{extension}$$

- Extension shifts binding to optimal fitness.

- Experimental extension ~ optimal.

- General design principle of molecular recognition systems?



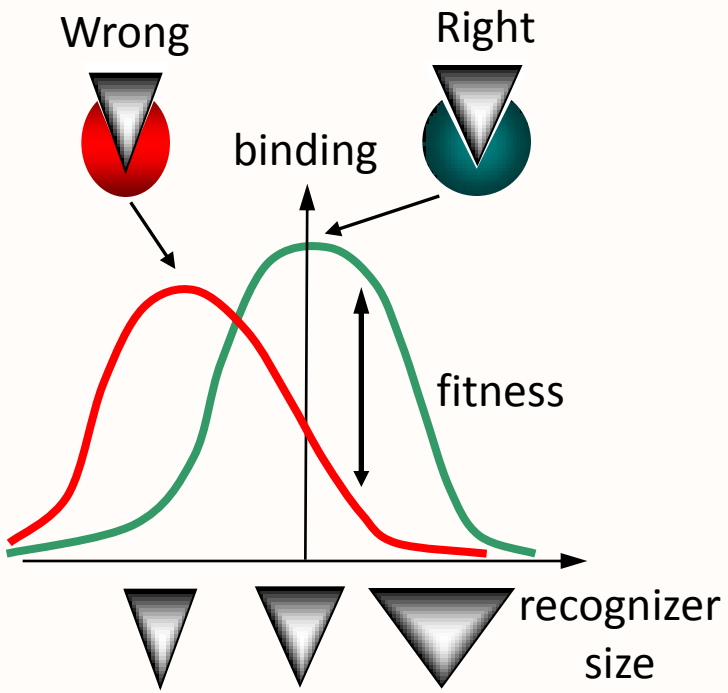
Optimal recognition: When off-target is right on

- Structural *mismatch or energy barrier*
- Reduce Right, *but* reduces Wrong even more.

- Result: Enhancement of fitness $\mathcal{F} \approx P_{right} - P_{wrong}$.

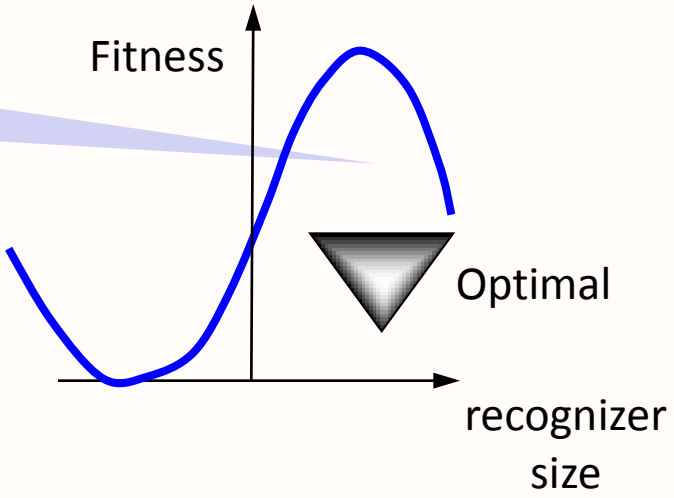
- **Conformational Proofreading:**

Optimal fitness at non-zero mismatch or extension.



- Optimal recognizer is *off-target*
- Not lock-and-key (*induced fit*)

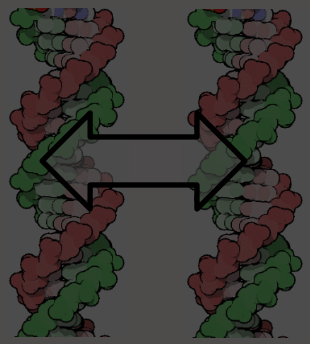
?



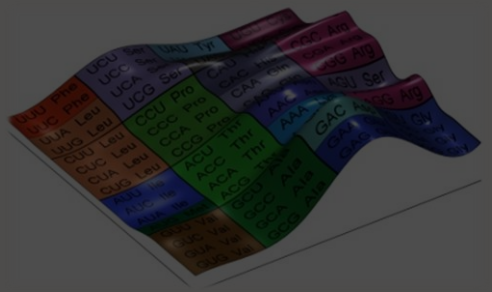
(Savir &TT, PLoS 1 2007, IEEE J 2008, Mol Cell 2010)

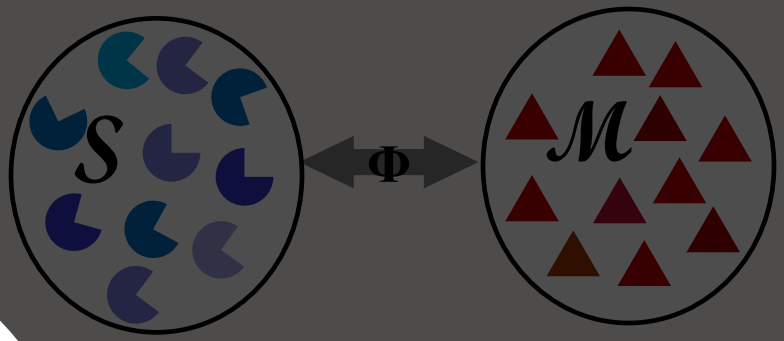
Examples: four essential living systems

Homologous Recombination

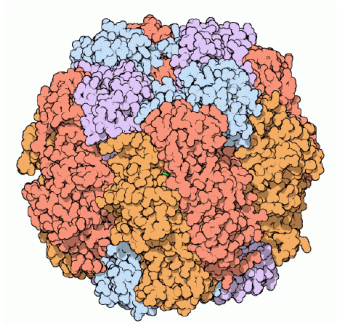


Molecular codes (genetic code)

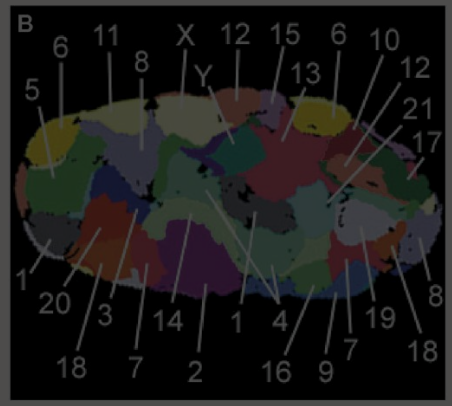




Photosynthesis

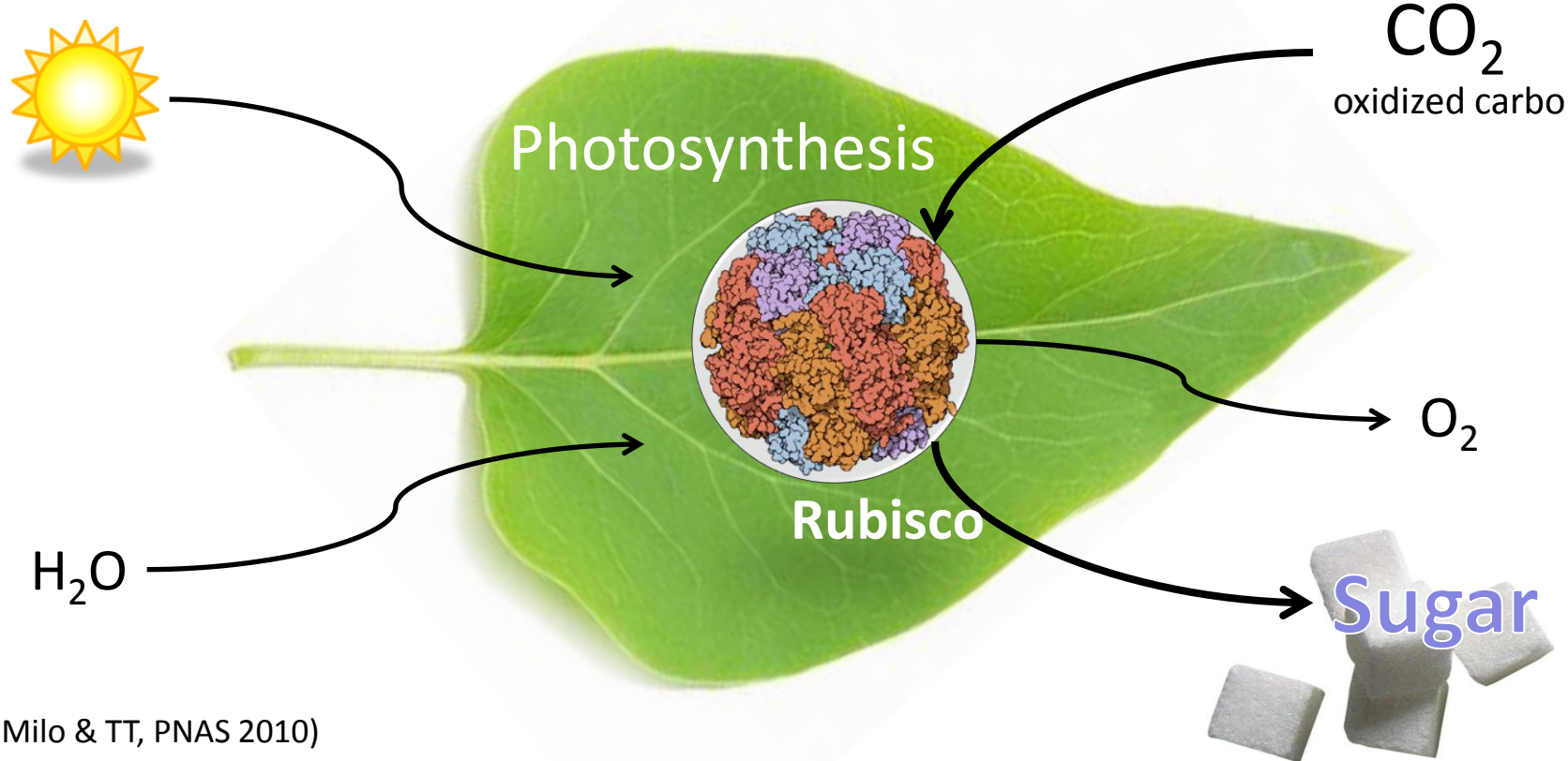


Chromosome organization



Rubisco captures atmospheric CO_2 to make sugar

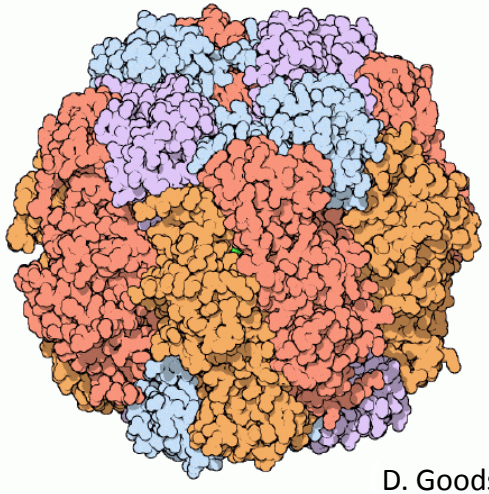
- Photosynthesis fixates carbon into organic forms.
- Rubisco captures CO_2.
- Complex of 8 large + 8 small subunits (540 kDa).



(Savir, Noor, Milo & TT, PNAS 2010)

Impact of Rubisco's (in)efficiency on the biosphere

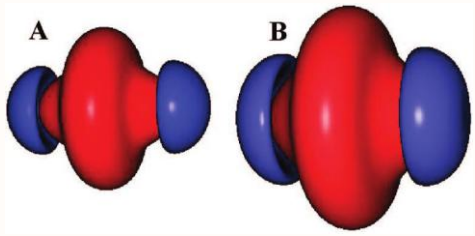
- Most abundant protein on Earth.
- Catalyzes most carbon fixation.



D. Goodsell (PDB)

- **Very slow** catalysis rate (~ 3-10 CO_2 /sec).
- **Low specificity:** confuses O=C=O and O=O.

- Can Rubisco be improved? – little success so far.



(Kannapan & Gready 2008)

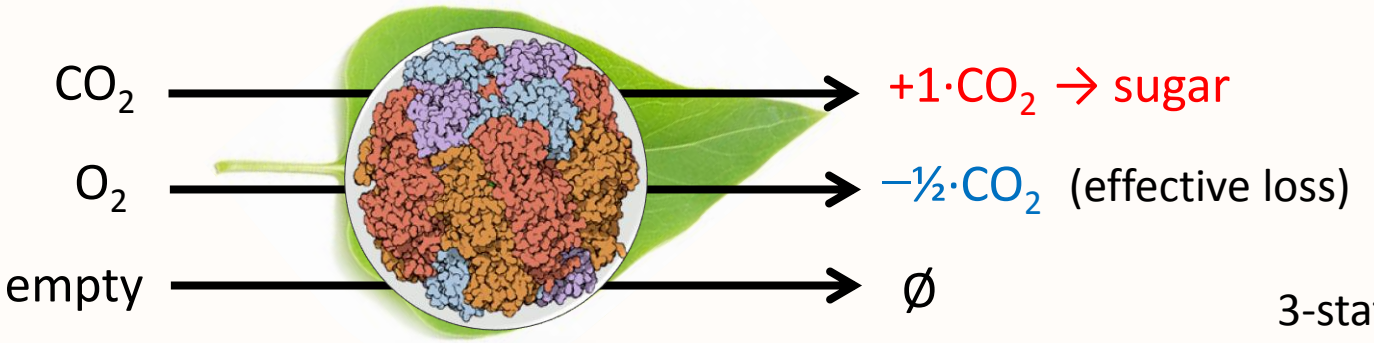
- Already optimized by evolution?
- Inefficiency due to biochemical constraints?

- Is Rubisco **optimal**, **constrained** or both?

Rubisco maps available substrates to products

- Four d.o.f.: 2 affinities to CO_2 and O_2 : $\{K_C \; K_O\}$

  2 catalysis rates : $\{\nu_C \; \nu_O\}$



- Overall fitness = net photosynthesis rate

$$\mathcal{F} = \sum_{\text{binding}} \frac{d}{dt}(\mathrm{CO_2})_b \times P_b = \nu_C P_{\mathrm{CO_2}} - \tfrac{1}{2}\nu_O P_{\mathrm{O_2}} + 0 \cdot P_{\text{empty}}$$

- Optimal Rubisco?

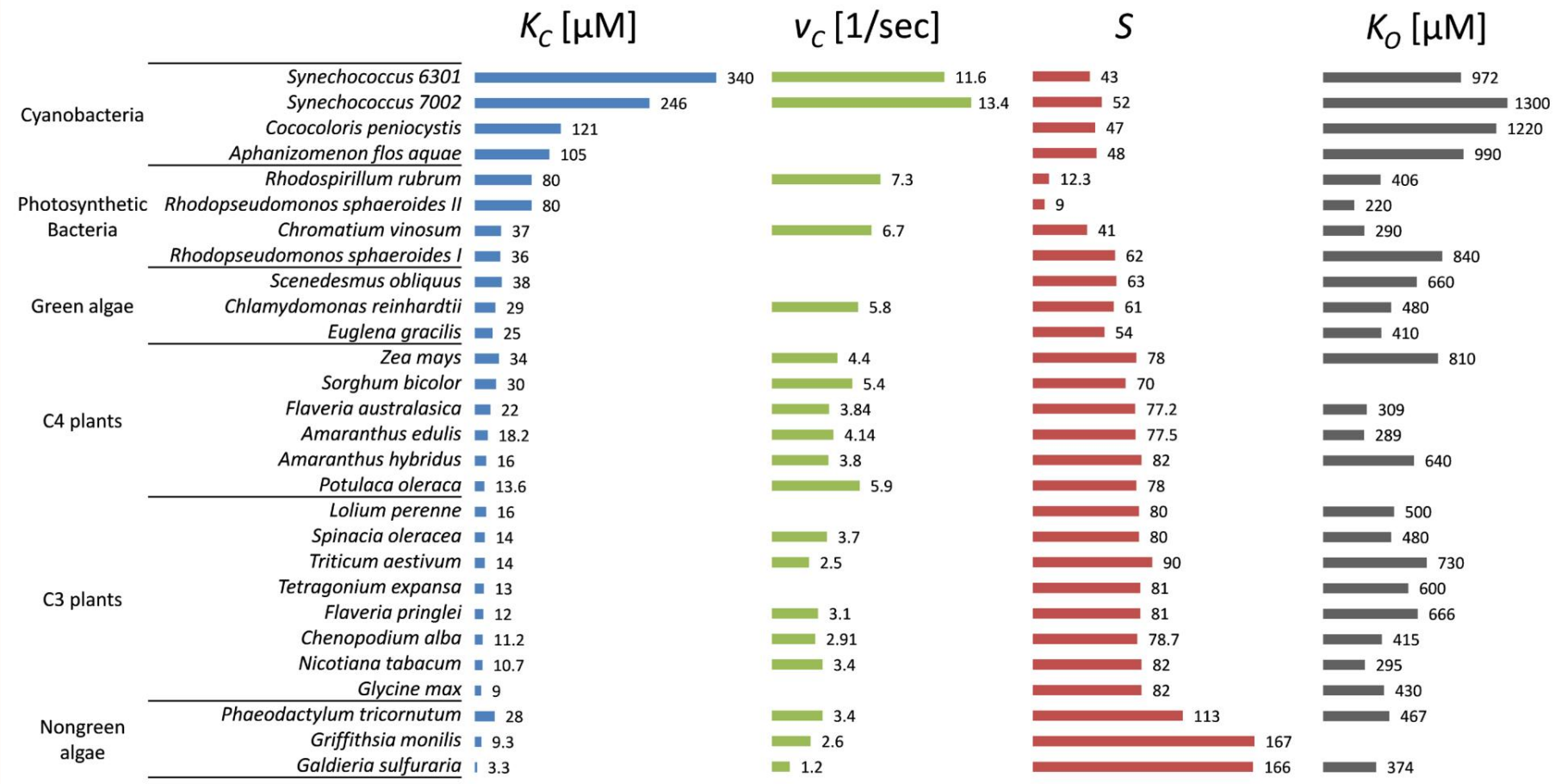
3-state partition

$$P_{CO_2} = \frac{\frac{[CO_2]}{K_C}}{1 + \frac{[CO_2]}{K_C} + \frac{[O_2]}{K_O}},$$

$$P_{O_2} = \frac{\frac{[O_2]}{K_C}}{1 + \frac{[CO_2]}{K_C} + \frac{[O_2]}{K_O}}.$$

Cross-species analysis exhibits strong correlation

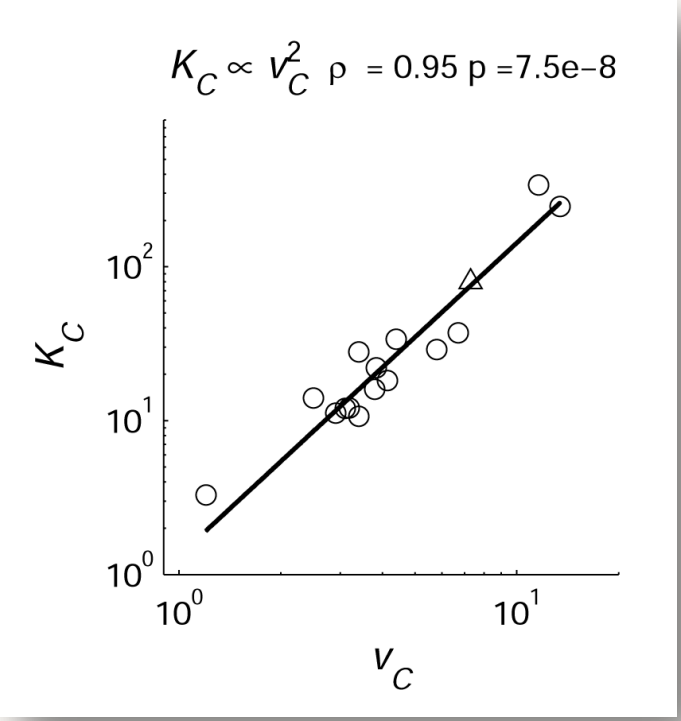
- Analysis of data from 28 eukaryotes and prokaryotes:



$$\text{Specificity,}\quad S = \frac{\text{carboxylation}/[\mathrm{CO_2}]}{\text{oxygenation}/[\mathrm{O_2}]} = \frac{v_C / K_C}{v_O / K_O}$$

Rubisco's parameter space is effectively 1D

- Data resides in **4D** space (S, v_C , K_C , K_O).
- But constrained to **1D power law** (linear in log scale).
- PCA analysis (>90% of variability is 1D) → power law correlations:



$$K_C \propto v_C^{2.0\pm0.2}$$
$$K_O \propto v_C^{0.5\pm0.1}$$
$$S \propto v_C^{-0.5\pm0.1}$$

Constrained Plasticity

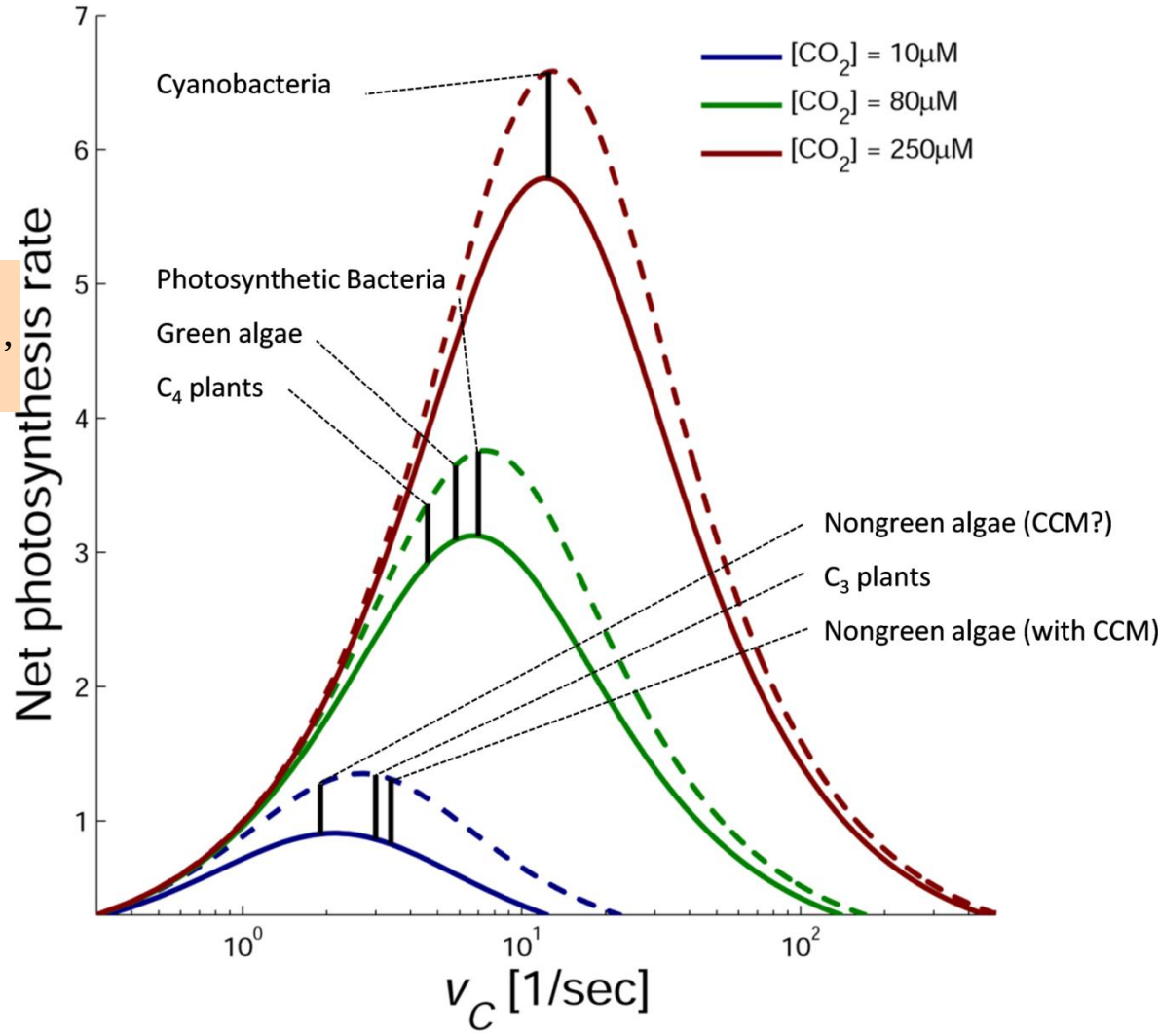
(Savir, Noor, Milo & TT, PNAS 2010)

Rubisco are nearly optimal to their habitats

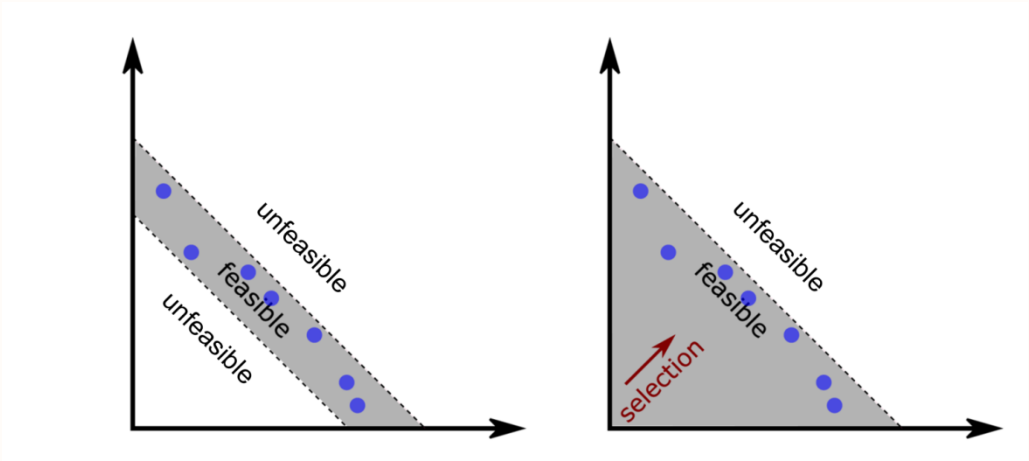
- Max($\mathcal{F}$) as function of v_C in given habitat (O_2 and CO_2).

$$\mathcal{F}(v_C) = \frac{v_C - 0.003 \cdot \frac{[O_2]}{[CO_2]} \cdot v_C^{3/2}}{1 + 1.3 \cdot \frac{1}{[CO_2]} \cdot v_C^2 + 0.005 \cdot \frac{[O_2]}{[CO_2]} \cdot v_C^{3/2}},$$

- All organism classes are nearly optimal.



Interplay of evolution and constraints shapes Rubisco

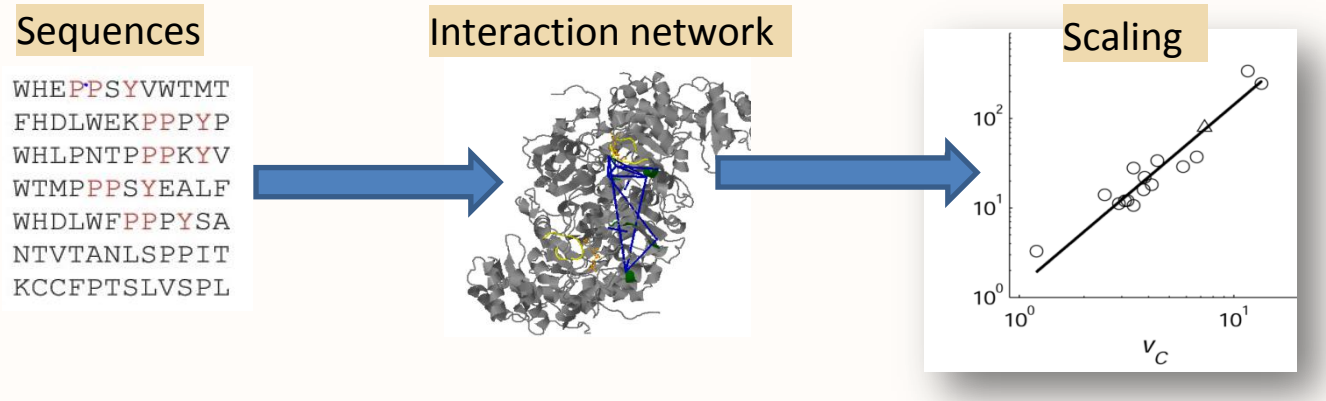


?

- Hint: stronger fluctuations for parameters that affect $\mathcal{F}$ weakly (K_O vs. K_C).
- Possible test: point mutation survey.

Questions, future directions, generalization…

- **Structural Mechanism?**



- Response to **long term climate changes**? (CO_2, Temp.…)

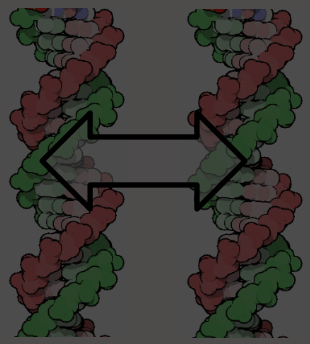
- **Constrained plasticity** in low dimensional landscapes:

  Generic phenomenon in proteins?

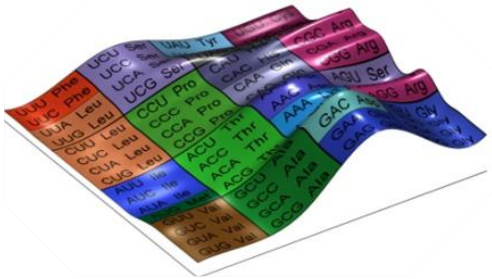
  – preliminary data from other strongly selected proteins.

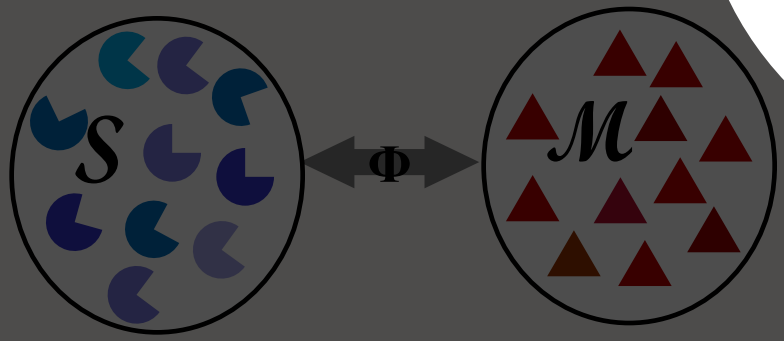
Examples: four essential living systems

Homologous Recombination

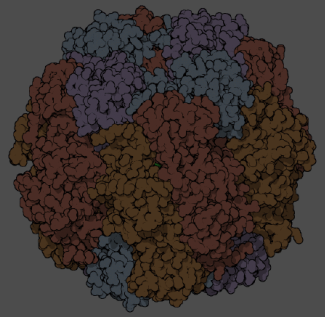


Molecular codes (genetic code)

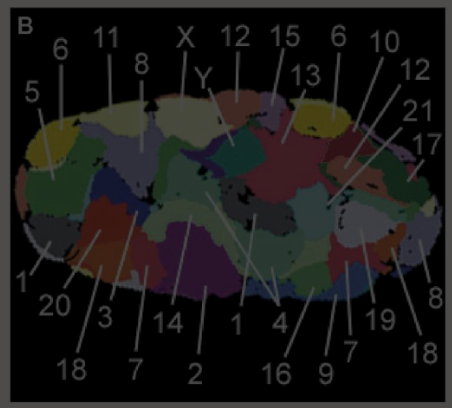




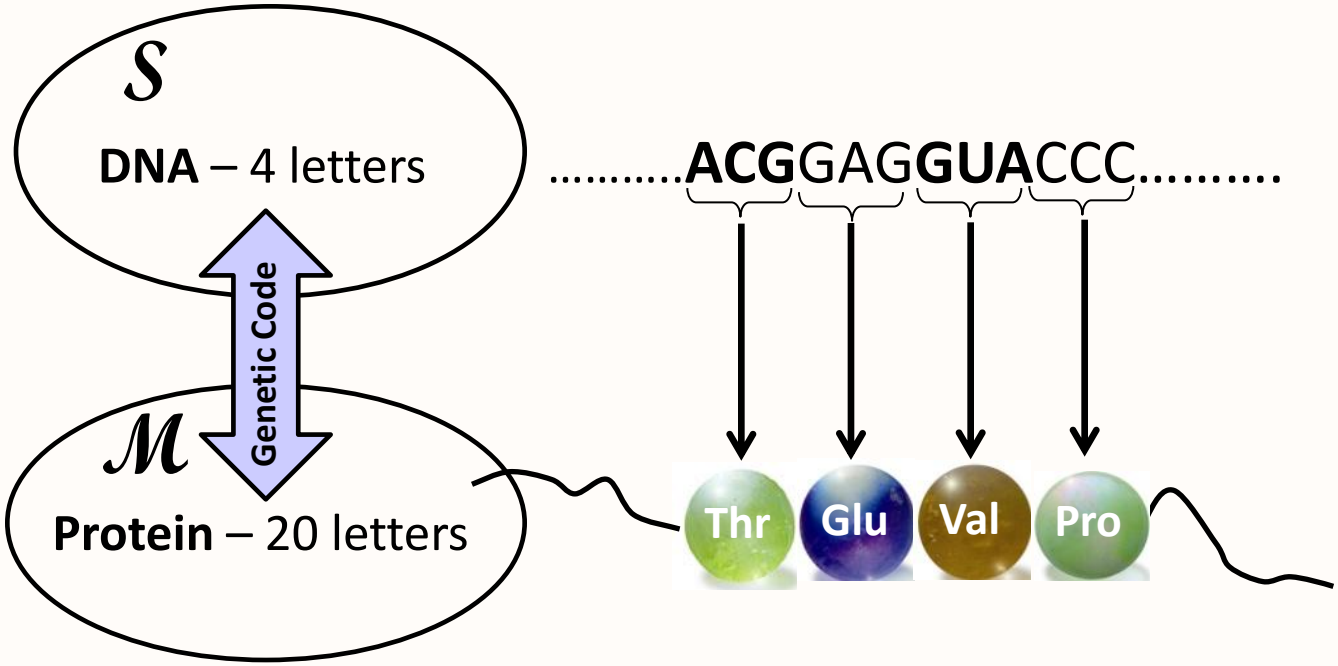
Photosynthesis



Chromosome organization



The genetic code is main info channel of life



- **Genetic code**: maps 3-letter words in 4-letter DNA language (4^3 = 64 codons) to protein language of 20 amino acids.

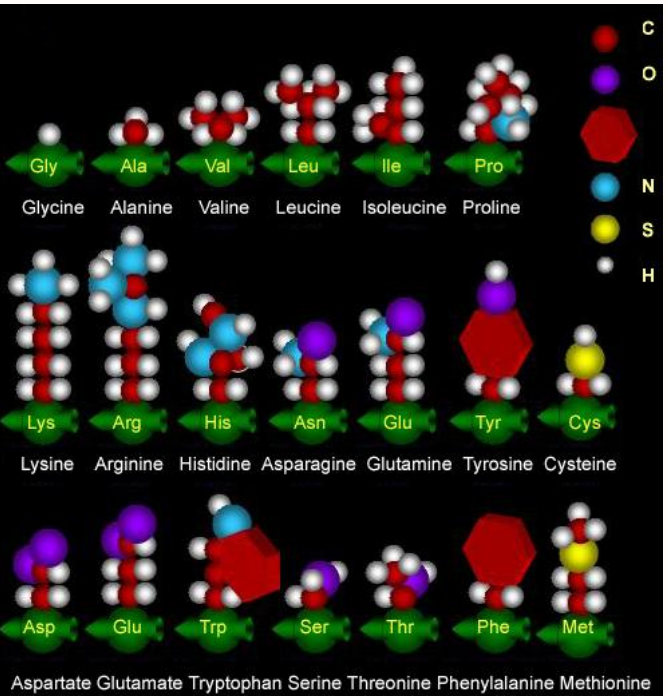
- Proteins are amino acid polymers.

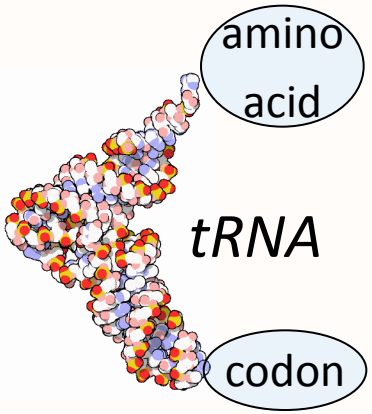
- **Diversity** of amino-acids is essential to protein functionality.

The genetic code maps codons to amino-acids

- Molecular code = map relating two sets of molecules (spaces, “languages”) **via molecular recognition**.
- Spaces defined by similarity of molecules (size, polarity etc.)

20 amino-acids





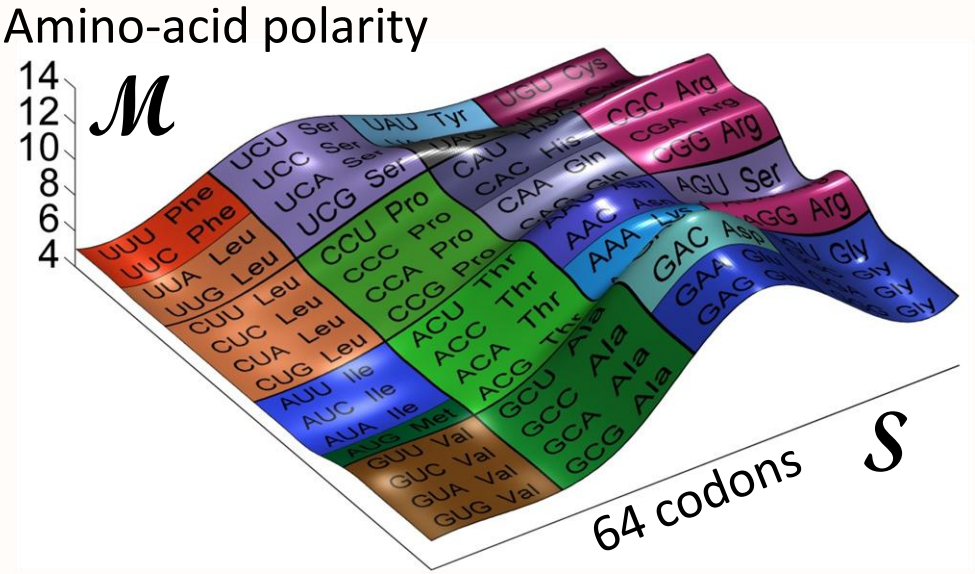
Genetic Code

64 codons

UUU	UUA	UCU	UCA	UAU	UAA	UGU	UGA
UUC	UUG	UCC	UCG	UAC	UAG	UGC	UGG
CUU	CUA	CCU	CCA	CAU	CAA	CGU	CGA
CUC	CUG	CCC	CCG	CAC	CAG	CGC	CGG
AUU	AUA	ACU	ACA	AAU	AAA	AGU	AGA
AUC	AUG	ACC	ACG	AAC	AAG	AGC	AGG
GUU	GUA	GCU	GCA	GAU	GAA	GGU	GGA
GUC	GUG	GCC	GCG	GAC	GAG	GGC	GGG

The genetic code is a smooth mapping

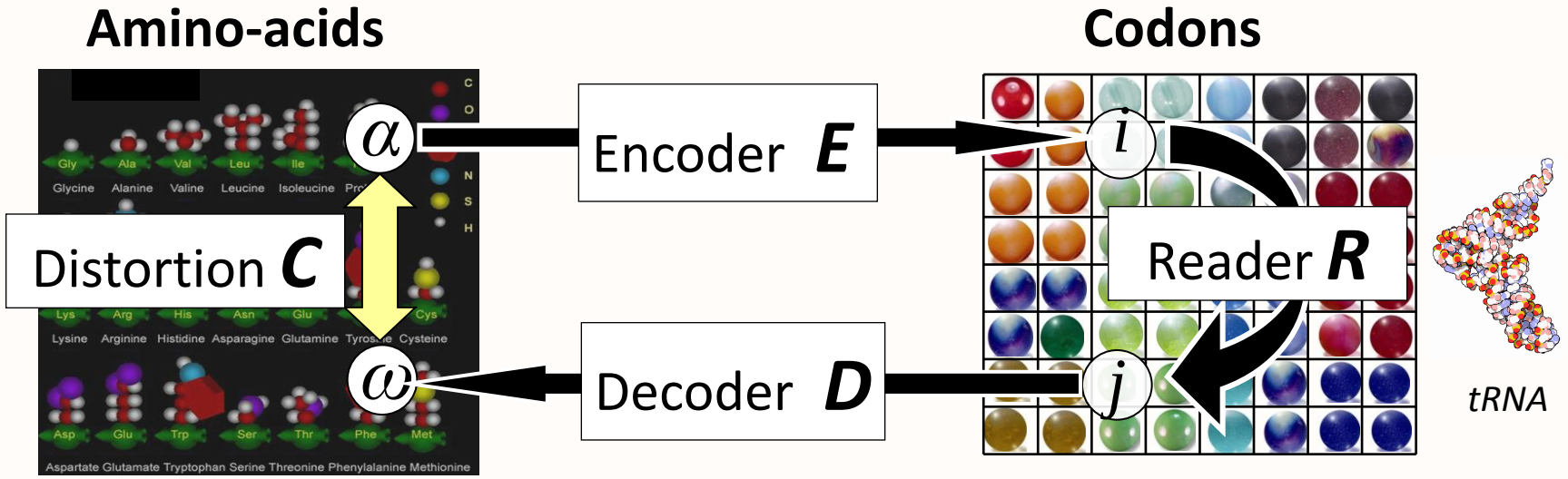
UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA TER	UGA TER
UUG Leu	UCG Ser	UAG TER	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly



- Degenerate (20 out of 64).
- Compactness of amino-acid regions.
- **Smooth** (similar “color” of neighbors).

Generic properties of molecular codes?

Fitter codes have minimal distortion

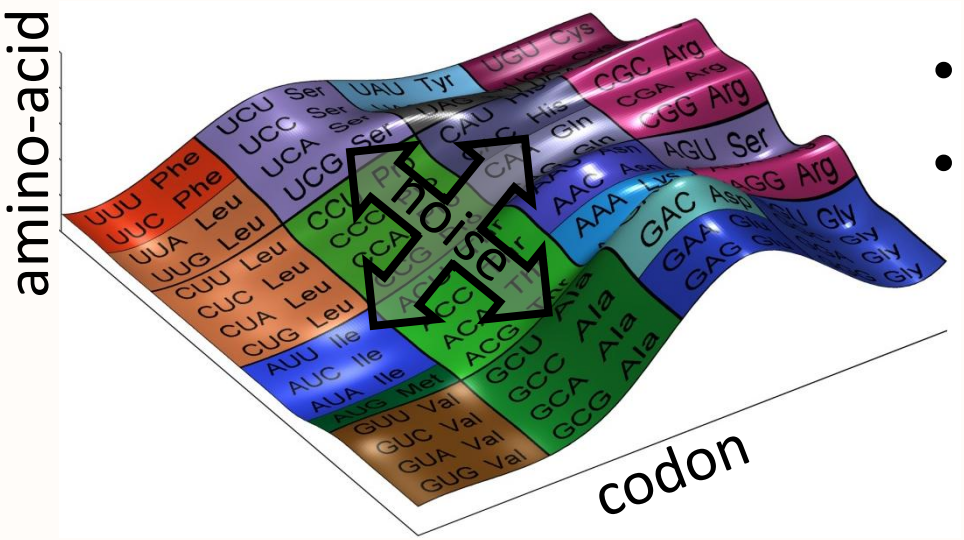


$$Q = \langle C_{\alpha\omega} \rangle = \sum_{paths} P_{path} C_{\alpha\omega} = \mathrm{Tr}\left(E \cdot R \cdot D \cdot C\right)$$

- Distortion of noisy channel, ***Q*** = average distortion of AA.
- ***R*** defines topology of codon space.
- ***C*** defines topology of amino-acid space.

(TT, J Th Bio 2007, PRL 2008, PNAS 2008)

Smooth codes minimize distortion

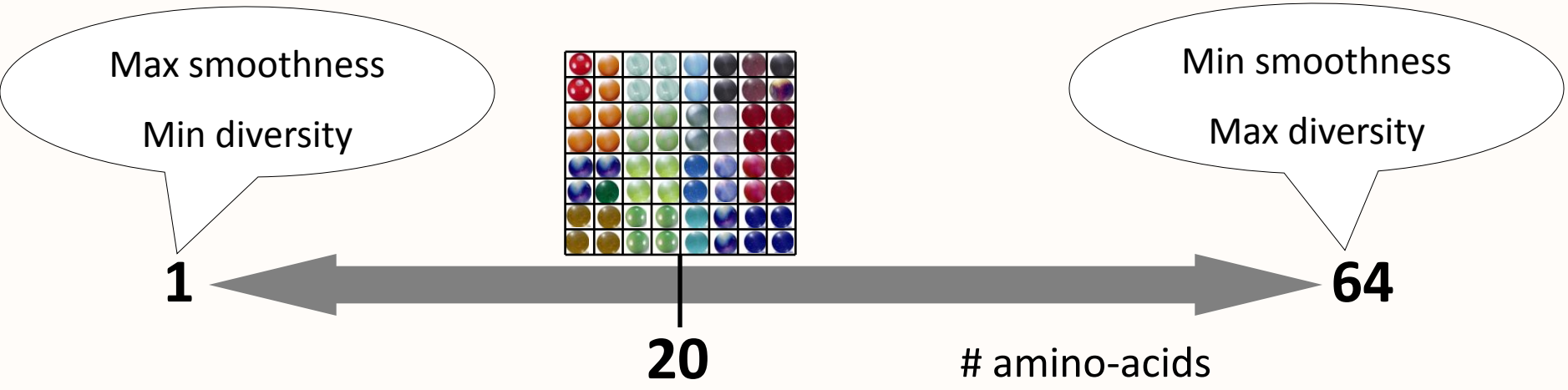


- Noise confuses close codons.
- Smooth code:

  close codons = close amino-acids.

  → minimal distortion.

- Optimal code must balance contradicting needs for **smoothness** and **diversity**.

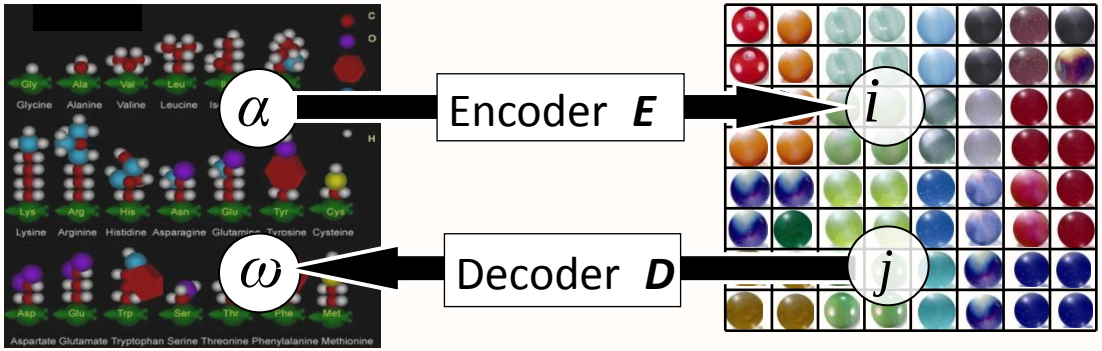


(TT, Bio Phys 2008)

Code's cost is the rate of the channel

- Diversity requires high **specificity** = high $\varepsilon_{\text{binding}}$.
- Cost ~ $< \varepsilon_{\text{binding}} >$.
- P_{binding} ~ Boltzmann: $E \sim e^{\varepsilon b / T}$.

$$I \sim \sum_{\alpha,i} E_{\alpha i} \ln E_{\alpha i} + \sum_{\alpha,i} D_{j\omega} \ln D_{j\omega} \sim \left\langle \varepsilon_{\alpha i} \right\rangle_E + \left\langle \varepsilon_{j\omega} \right\rangle_D$$



- Cost ***I*** = **Channel Rate** (bits/message)

Code fitness combines rate and distortion of map

Fitness = −Gain x *Distortion − Rate:*

"metrics"

$$\mathscr{F}(\beta, E, D, R, C) = -\beta Q - I$$

maps

- Fitness $\mathscr{F}$ is "free energy" with inverse "temperature" β.
- ***Gain*** β increases with organism complexity and environment richness.
- Evolution varies the gain β.

- Population of self-replicators evolving according to code fitness $\mathscr{F}$: mutation, selection, random drift.

(TT, PRL, Bio Phys 2008)

Code emerges at a critical coding transition

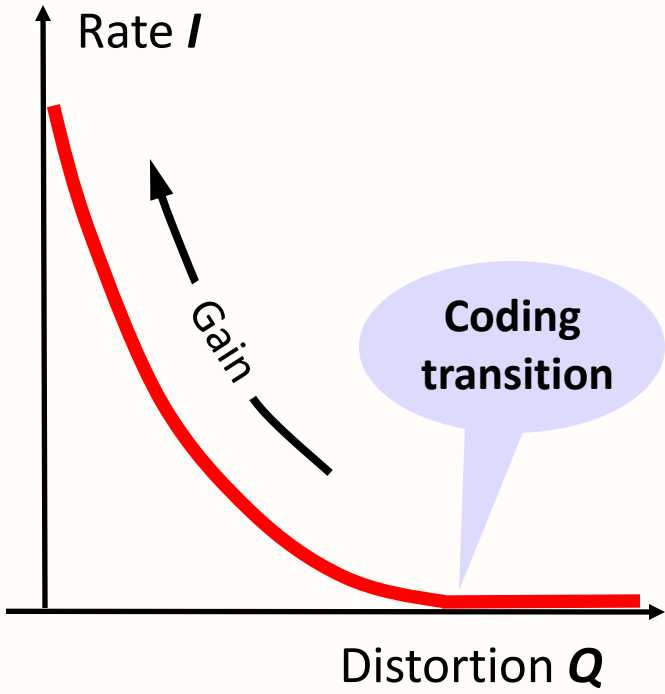
- Low gain β : Cost too high

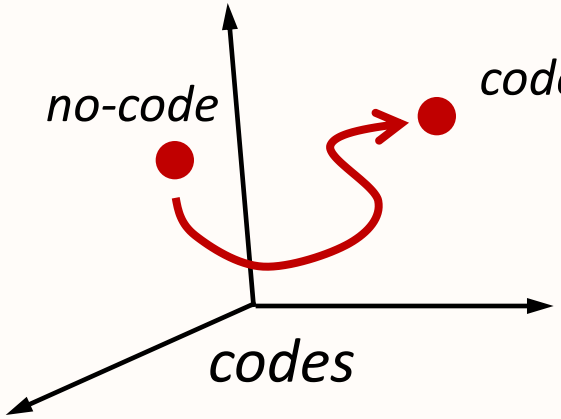
  → no specificity → **no code**.

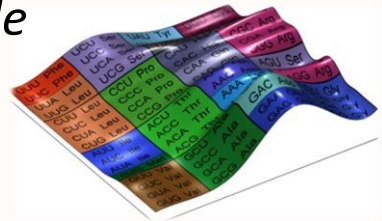
- β increases: **Code emerges**

  channel starts to convey information ($I \neq 0$).

- Continuous 2nd order phase transition.
- Emergent map is smooth, low mode of ***R***.







Rate-distortion theory (Shannon 1956)

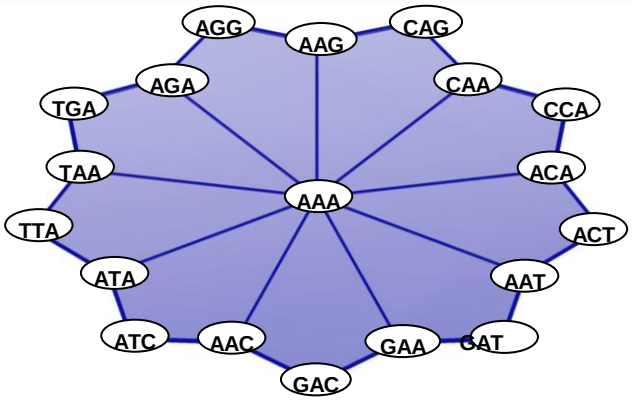
Errors define the topology of the genetic code

- Codon graph = codon vertices + 1-letter difference edges (mutations).

$K_4 \times K_4 \times K_4$

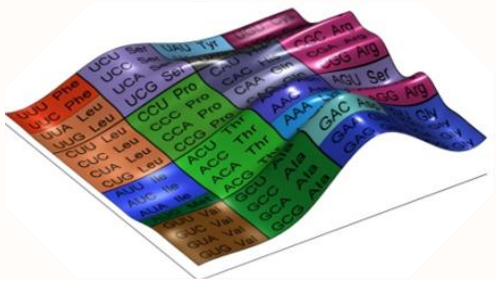
- Lowest excited modes of graph-Laplacian R .
- Single maximum for lowest excited modes (Courant).
- Every mode corresponds to amino-acid :

 # low modes = # amino-acids.



→ single contiguous domain for each amino-acid.

→ **Smoothness**.



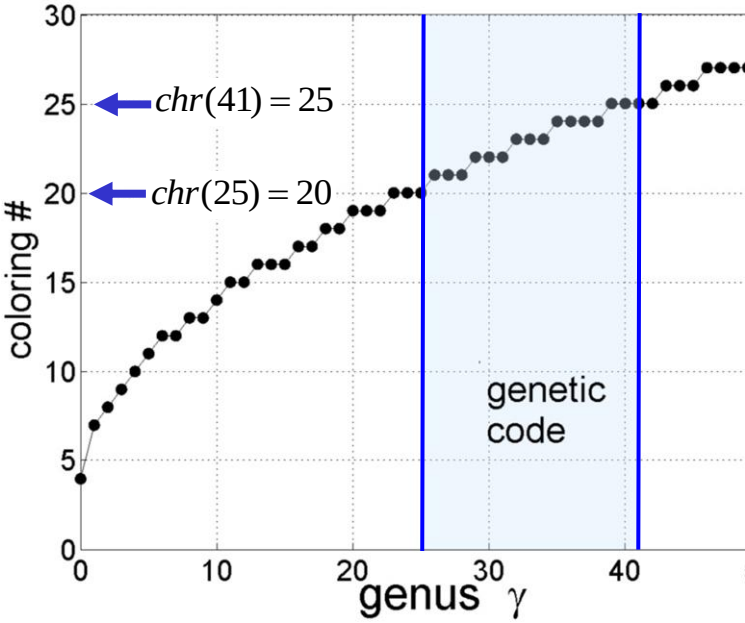
Coloring number limits number of amino-acids

- Coloring number = Min(# colors that suffices to color a map)
- Topological invariant (function of genus):

$$chr(\gamma) = \left\lfloor \frac{1}{2}\left(7 + \sqrt{1+48\gamma}\right) \right\rfloor .$$

(Ringel & Youngs 1968)

$$\max(\#\ \text{amino-acids}) = chr(\gamma)$$

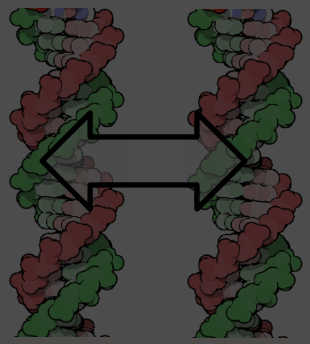


- From Courant 's theorem + "convexity" (tightness).
- Genetic code: γ = 25-41 → coloring number = 20-25 amino-acids

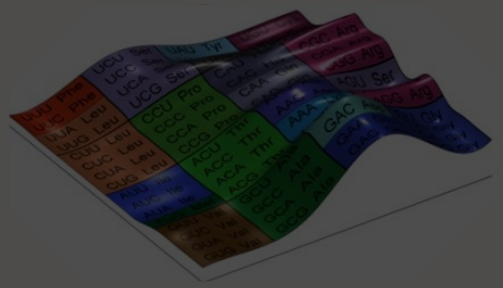
(TT, J Th Bio 2007, PRL 2008, PNAS 2008, *J Lin Alg* 2008, Phys Life Rev 2010)

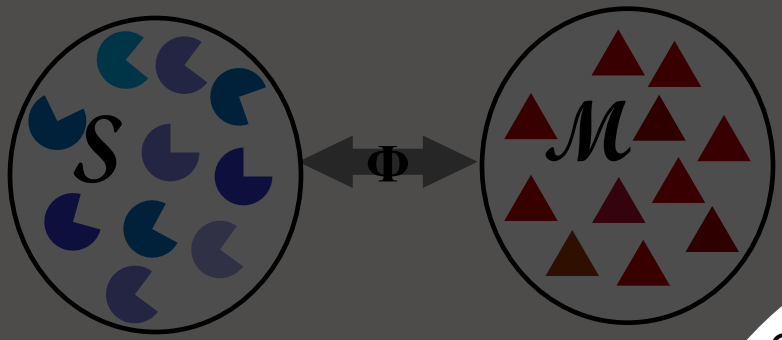
Examples: four essential living systems

Homologous Recombination

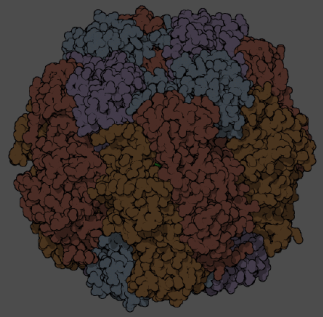


Molecular codes (genetic code)

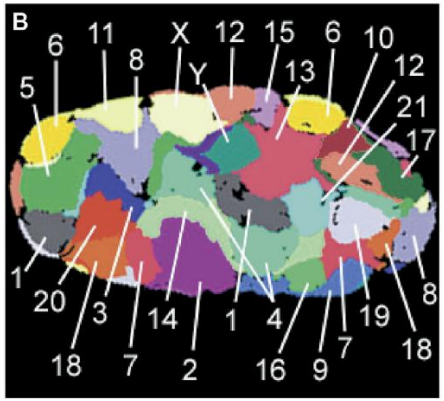




Photosynthesis



Chromosome organization



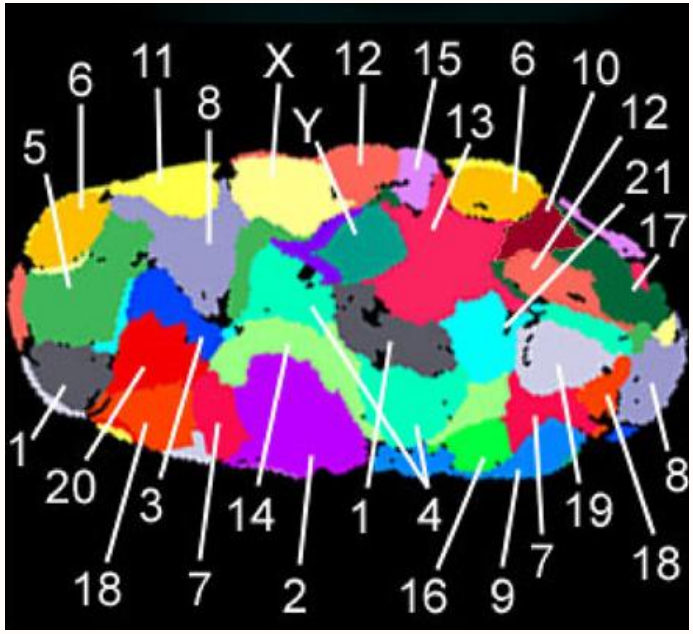
A Libchaber (Rockefeller)

JP Eckmann (Geneve)

GV Shivashankar (NUS)

The problem of chromosomes

- Genes are divided into chromosomes.

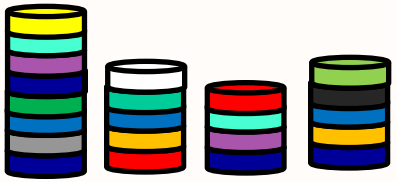


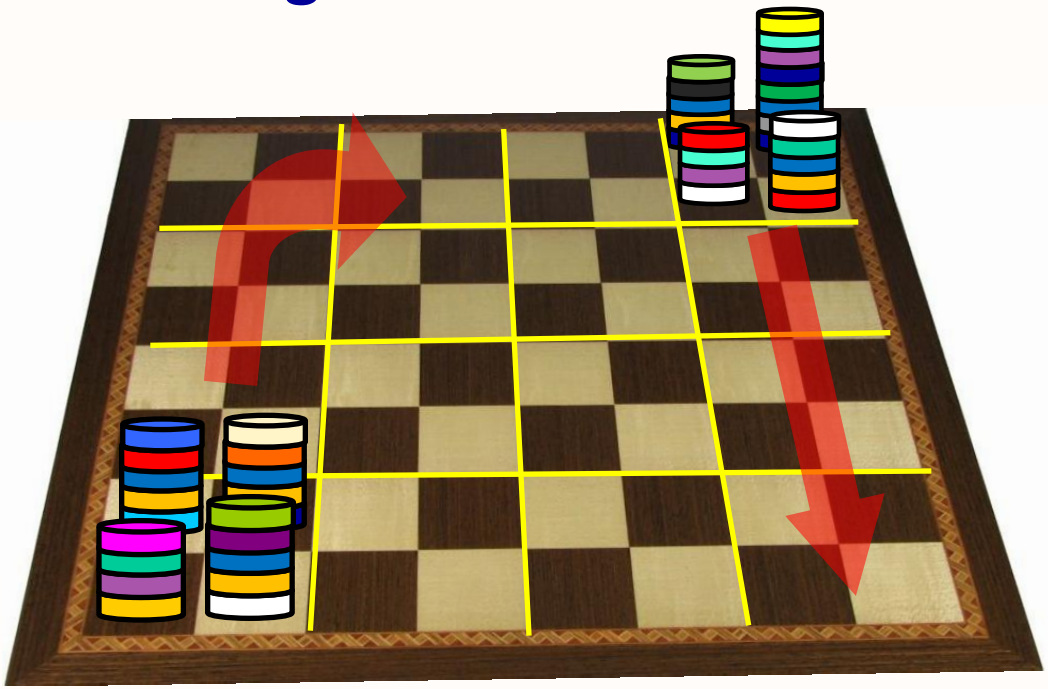
Nucleus of human fibroblast (Cremer et al., 2005)

- What is the optimal (?) number?
- What is the optimal (?) organization?
- Relation to cell type?

Organism	# chr
M. pilosula (ant)	2
Fruit fly	8
Arabisdopsis	10
C. elegans	12
Rye	14
Corn	20
Chinese hamster	22
Budding yeast	32
Earthworm	36
Cat	38
Syrian Hamster	44
Human	46
Tobacco	48
Silkworm	56
Horse	64
Dog	78
Goldfish	100
Adder’s tongue	1400

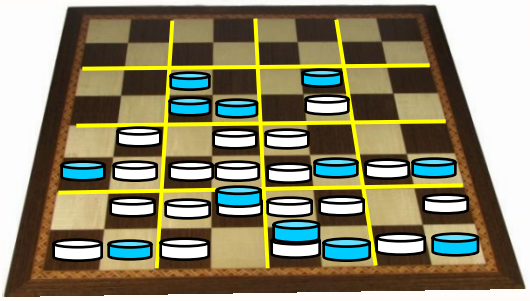
The Game of Chromosome Organization

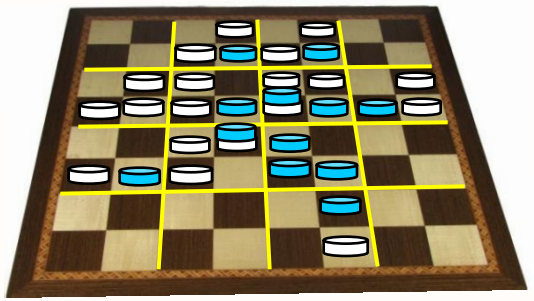




"Casino chip shuffling"

(A) Given stacks of multi-color Chips.
(B) Divide chessboard into "Blocks".
(C) Cover board with stacks.
(D) For each color: Shuffle the Blocks such that all chips of this color will be close as possible to each other.





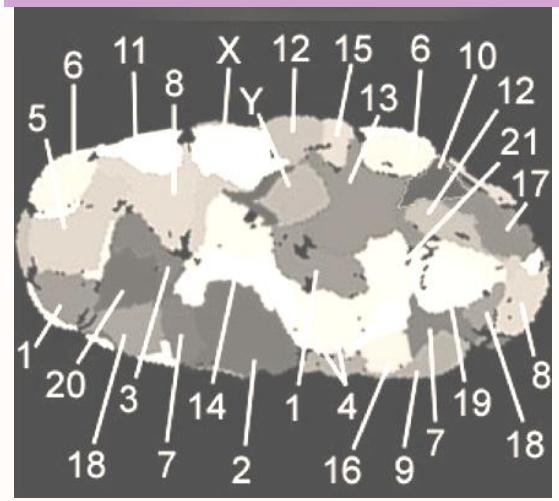
Mapping between 3D organization and expression

$\mathcal{S}$
Real Space

Cell Types

$\mathcal{M}$
Function Space

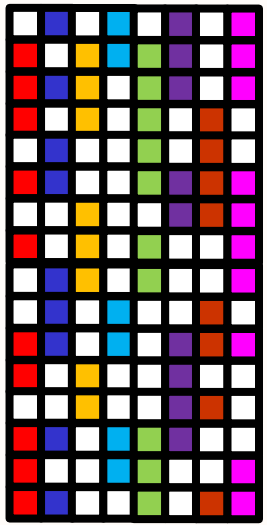
Spatial organization



Expression

Cell type

genes



?

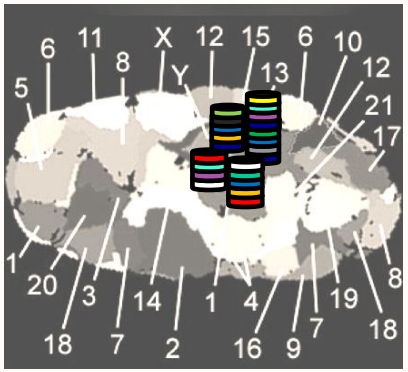
Hypothesis: relation between real space and function space based on optimality

Optimal chromosome organization?

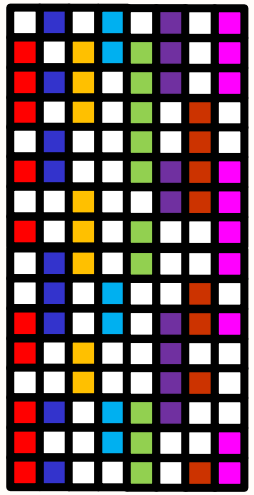
Hypothesis: co-expressed genes or active genes with similar function tend to reside in the same or in close chromosomes.

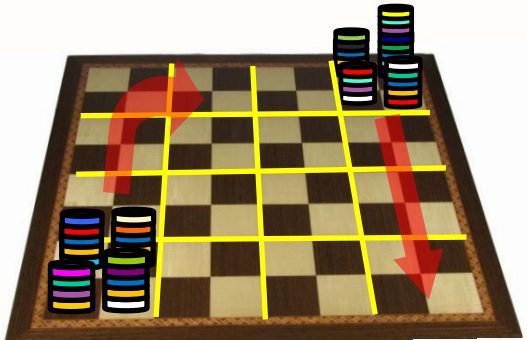
Optimal organization	Casino chip shuffling
Cell type/function	color
Gene	Chip stack
Chromosome	Block
Nucleus	Chessboard
Chrom. reorganization	Block shuffling
Close active genes	Close Same color chips

Nucleus



Expression





Possible mechanisms of chromosomal interactions

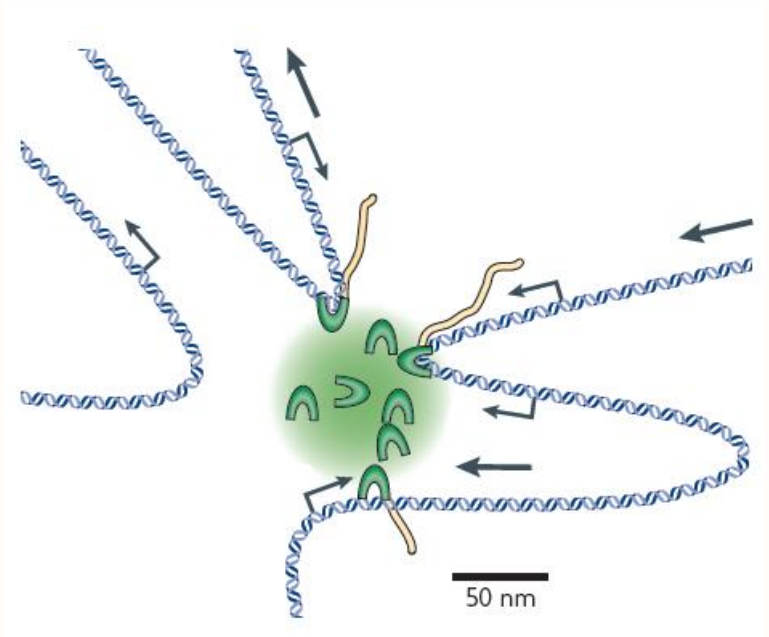
Expression space

- Genetic networks, co-regulation, co-expression.

Physical space

- Physical proximity boosts efficiency of transcription factories?
- Small nuclear RNA (snRNA)?

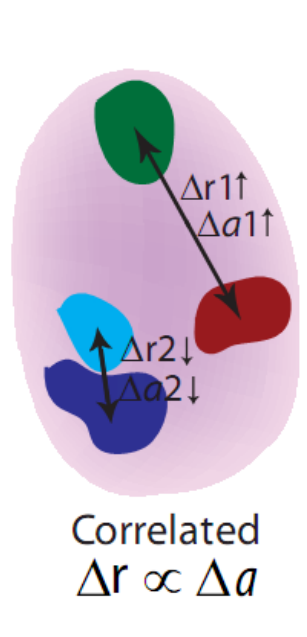
  small nuclear ribonucleoproteins (snRNP)?
- ….
- **Smoothness**

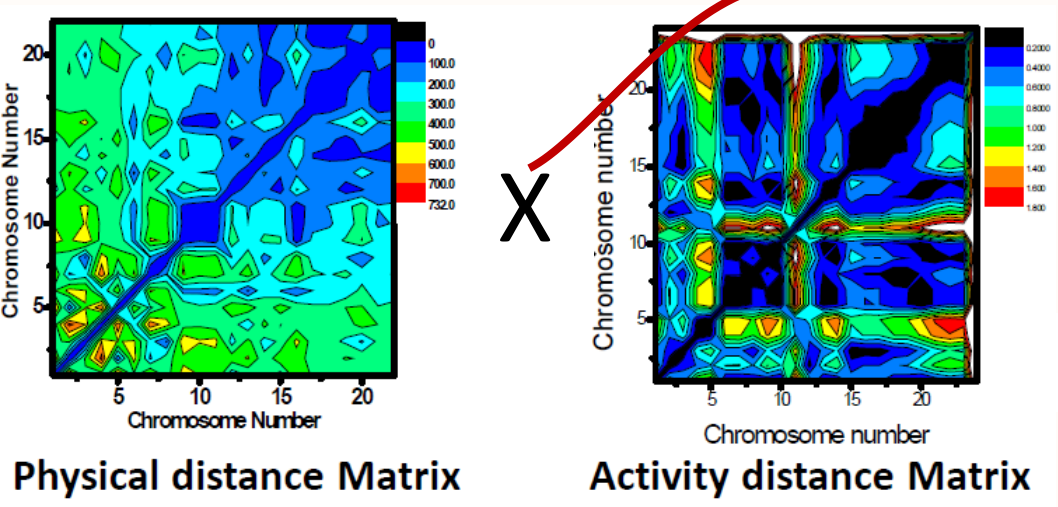


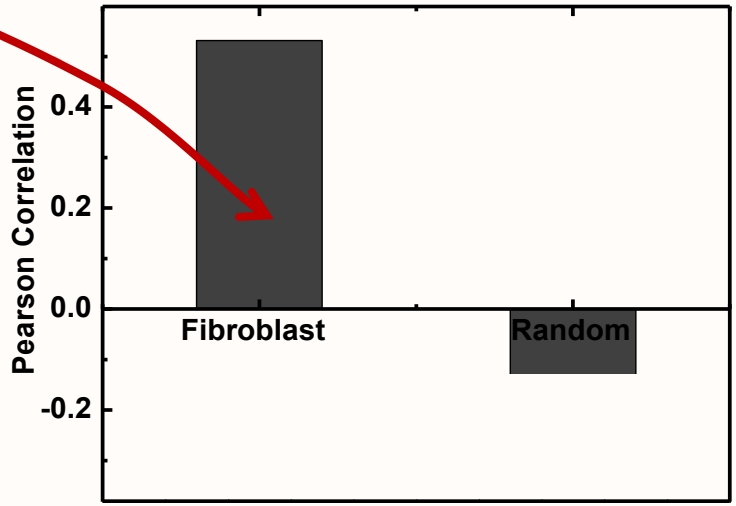
Transcription factories: Genes from same or from different chromosomes may associate with polymerases in the same factory. (Sutherland & Bickmore, *Nat Rev Gen* 2009)

Spatial organization and expression are correlated

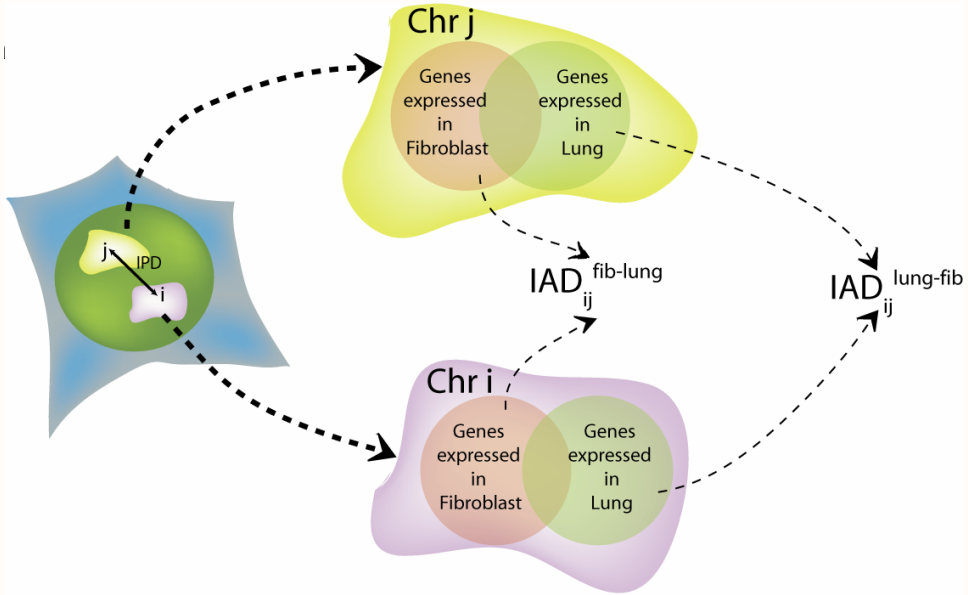
- Expression distance $\left| \ln \frac{\langle \text{activity/gene} \rangle_i}{\langle \text{activity/gene} \rangle_j} \right|$
- Physical distance $\left| \langle r_i - r_j \rangle \right|$
- Average over 54 nuclei yields significant correlation.

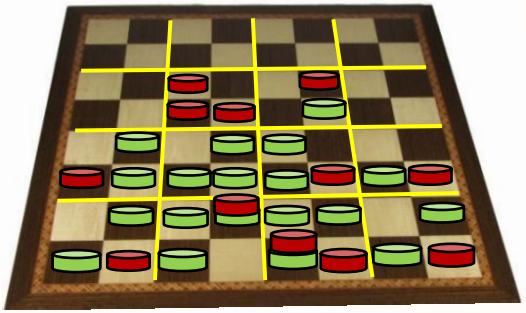


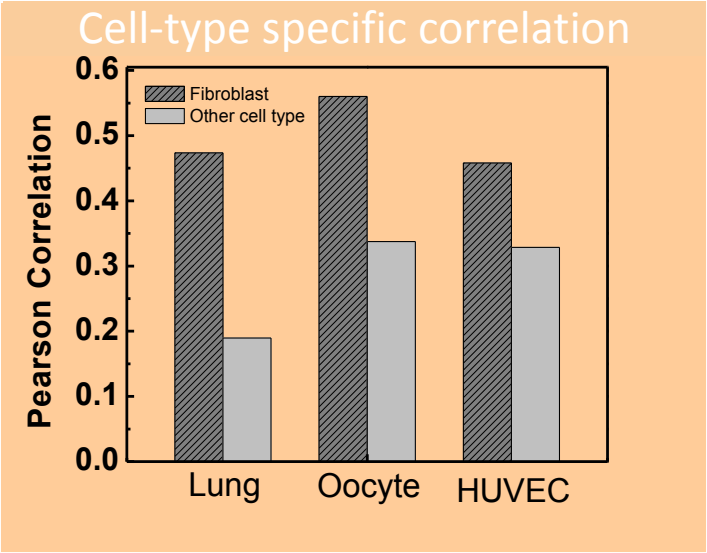




Is chromosome organization cell specific?

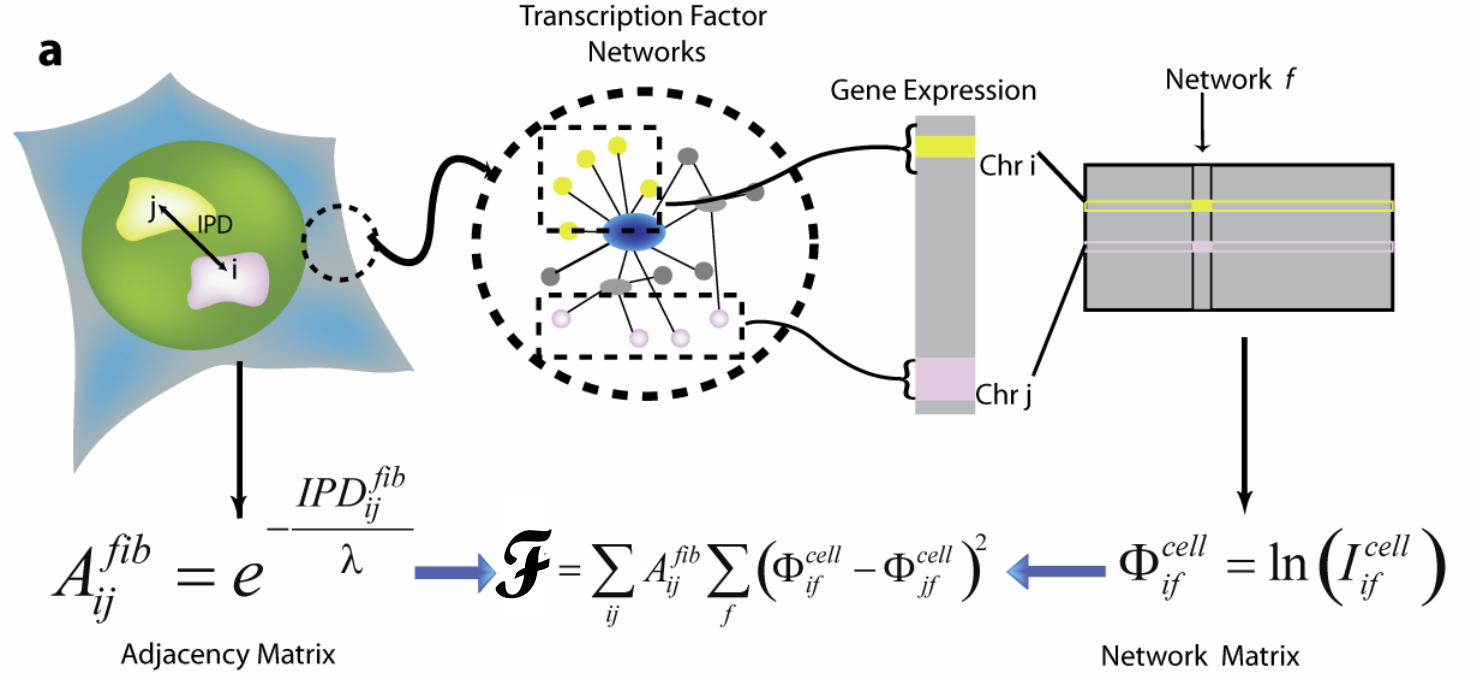






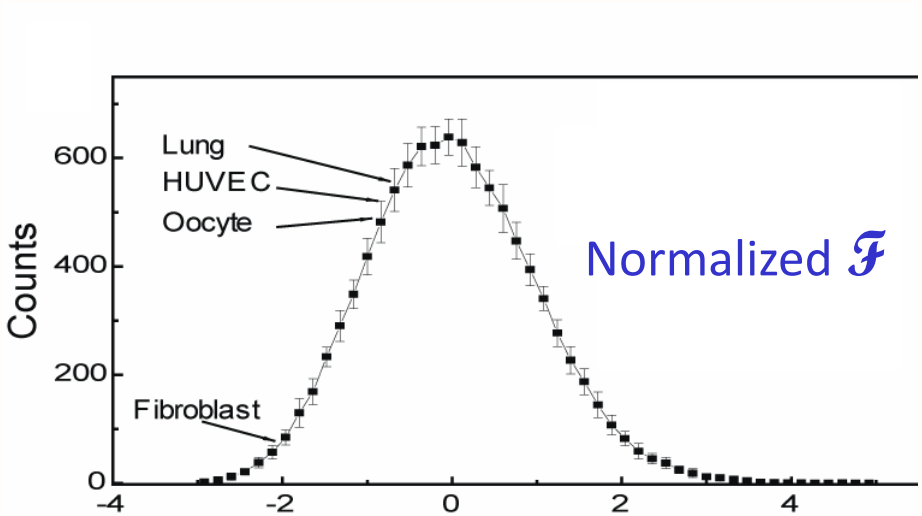
- HUVEC - human umbilical cord vein endothelial cell.
- Oocyte – female germ cells.
- Fibroblast – connective tissue.
- Lung – epithelial cells.

“Fitness” of chromosome organization



- $\mathcal{F}$ is “Smoothness” measure (~ distortion *Q*):

  How close are gene that perform the same function in given cell type?



Questions and directions

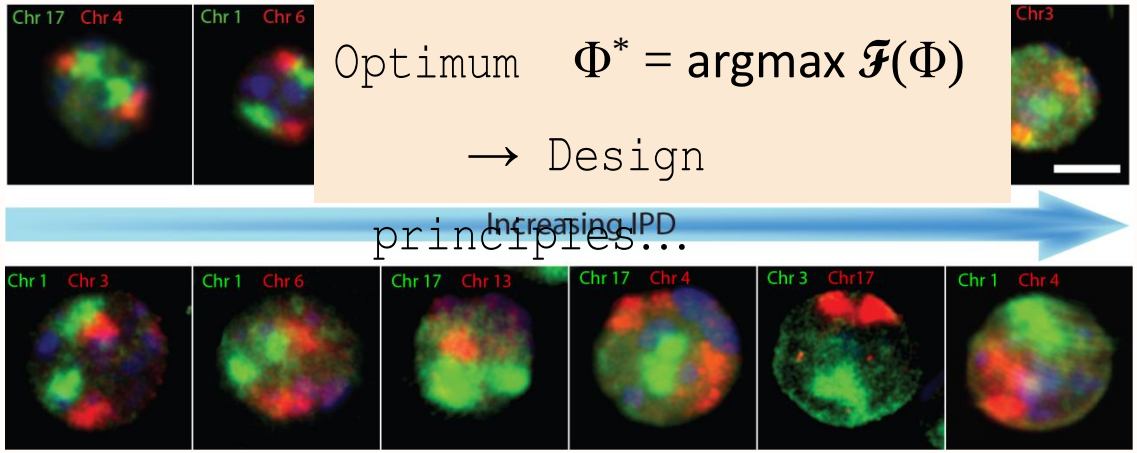
- Better optimality measures (transcription factories).
- Other cell types (preliminary evidence from T cells).
- Optimality tra… …ome #.
- Relations to th… …graph (A_{ij}).

Maps $\Phi: \mathcal{S} \leftrightarrow \mathcal{M}$

Fitness $\mathcal{F}(\Phi)$

Optimum $\Phi^* = \text{argmax}\, \mathcal{F}(\Phi)$

→ Design principles…



- Other molecular information channels:

molecular recognition, transcription networks.

THANKS

WIS

Yoni Savir

Uri Alon

Elisha Moses

Shalev Itzkovitz

Guy Shinar

Roy Bar-Ziv

Tsevi Beatus

Elad Noor

Ron Milo

Rockefeller

Albert Libchaber

NUS, NCBS

G.V. Shivashankar

S. Maharana

V.K. Iyer

Geneva

Jean-Pierre Eckmann

more: www.weizmann.ac.il\complex\tlusty