

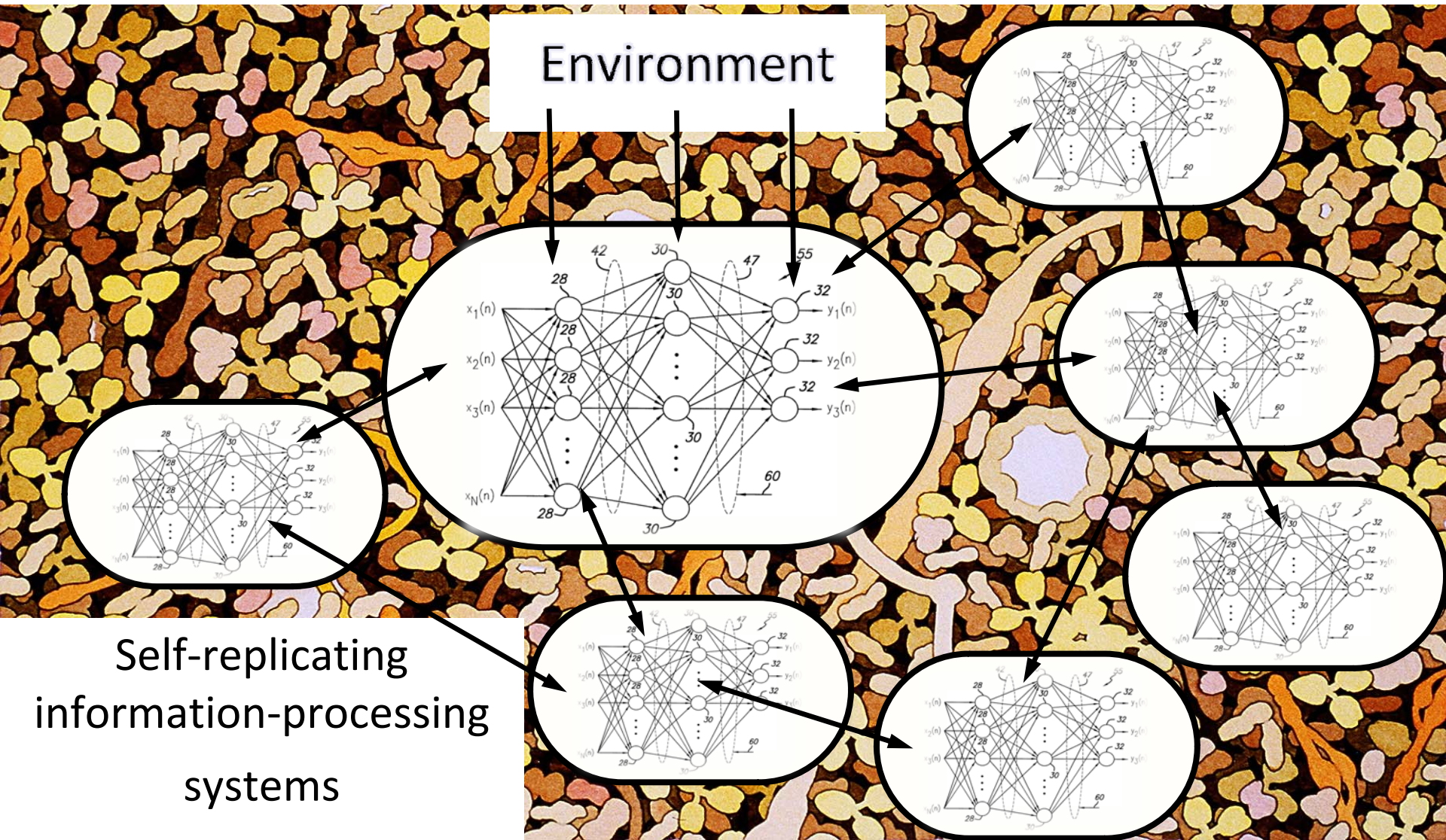
The Physical Language of Molecules

How do molecular codes emerge and evolve?

International Workshop on Bio-Soft Matter

Tokyo, 2008

Biological information is carried by molecules



Outline

- Molecular codes = **information channels** or **maps**.
- **Fitness** of codes = **Quality** + **Cost**.
- **Smooth** molecular codes emerge at **phase transitions**.
- **Topology** of errors governs emergent code.
- **Evolutionary dynamics** of codes.

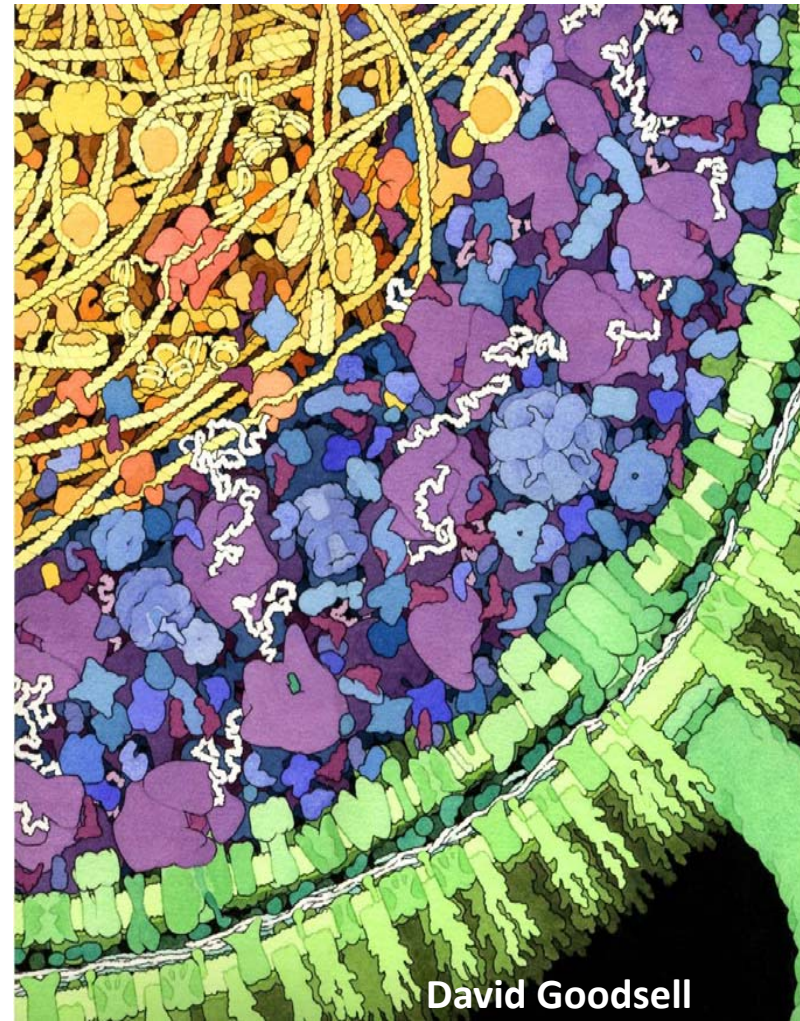
Challenge of molecular coding

Quality:

- Information transfer via molecular recognition in a noisy, crowded milieu.
- Recognizer and target fluctuate.
- Many competing *lookalikes*.
- Weak recognition interactions $\sim k_B T$.

Cost:

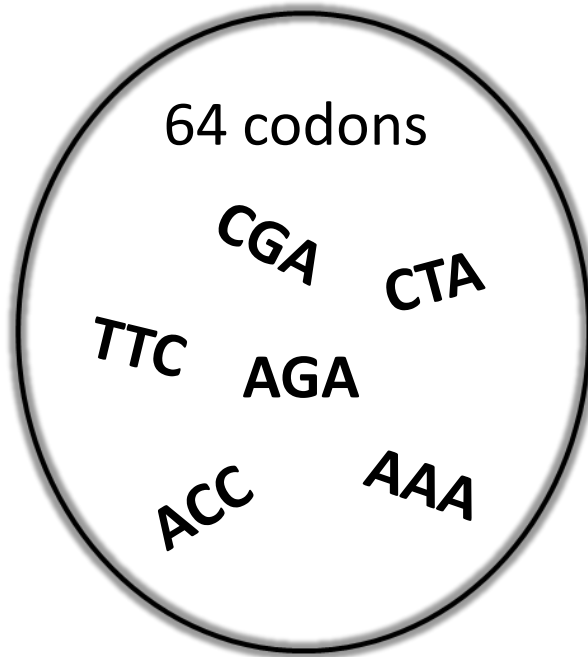
- How to construct the molecular codes at minimal cost of resources?



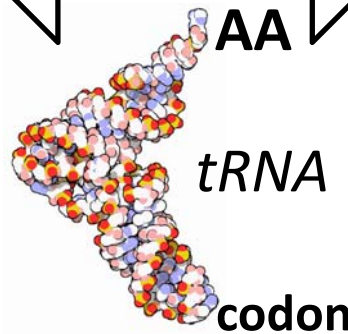
Coding theory: Molecular Codes as Maps or Information Channels

- Molecular code = *map* relating two sets of molecules.
- Relation by *molecular recognition*.

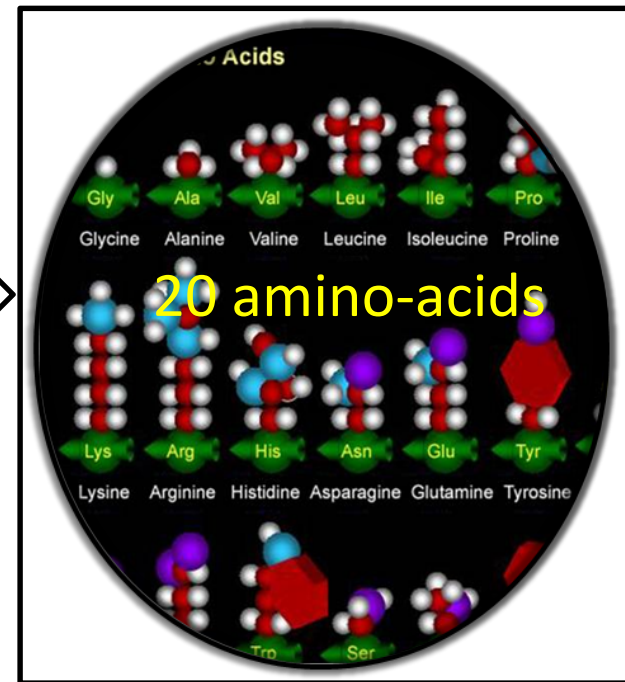
Symbols



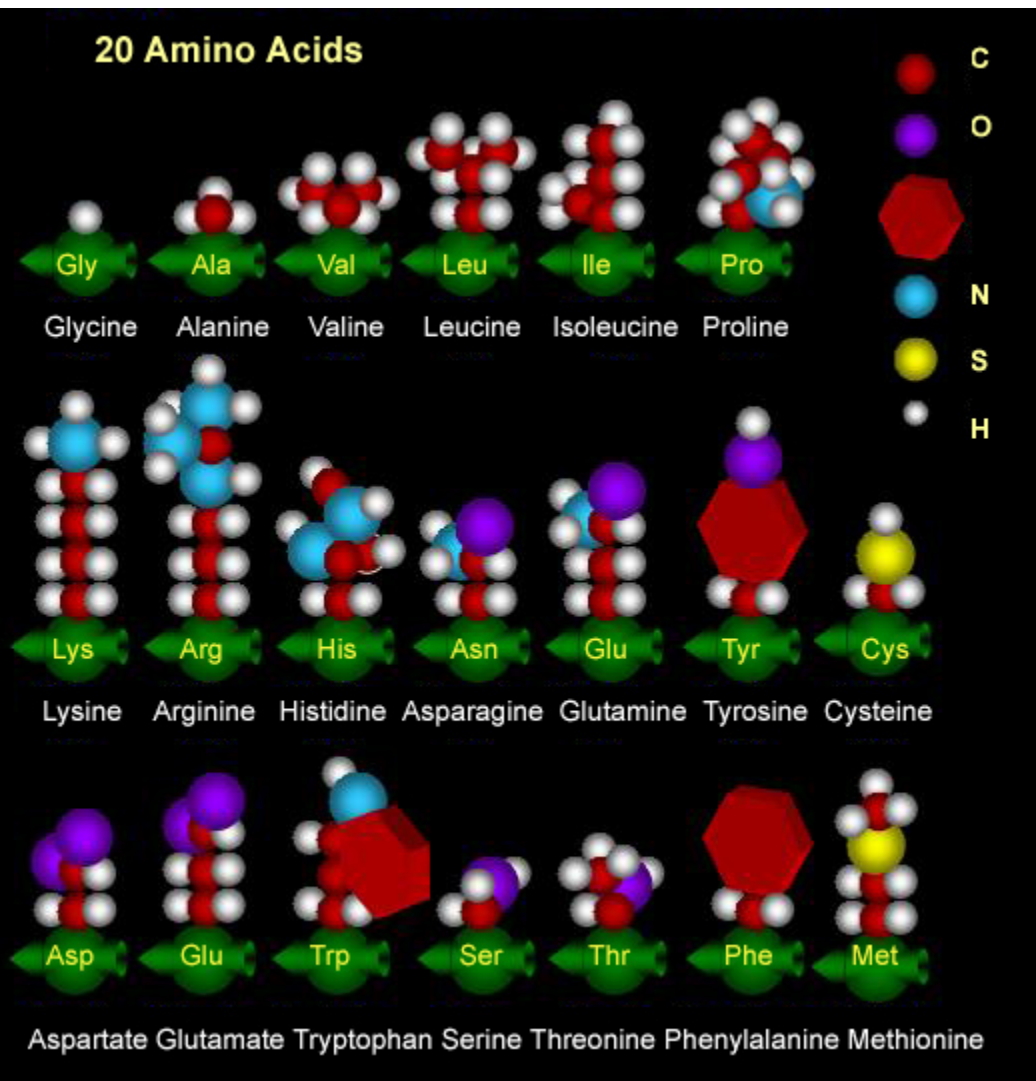
Genetic Code



Meanings



Amino acids are the building blocks of proteins

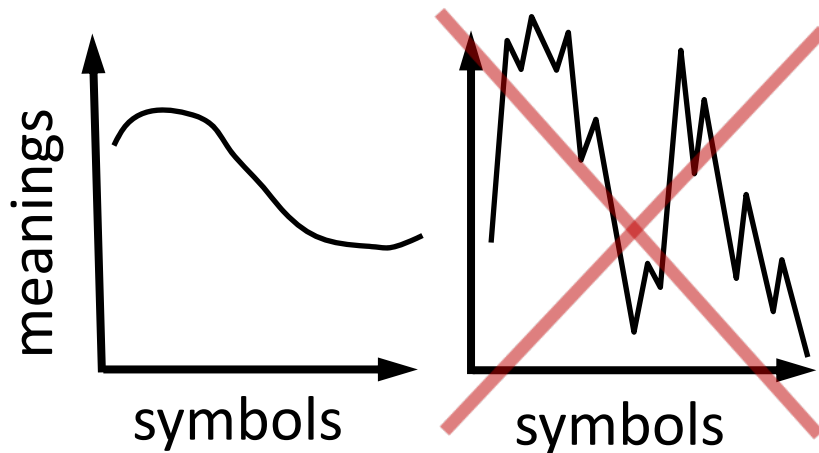
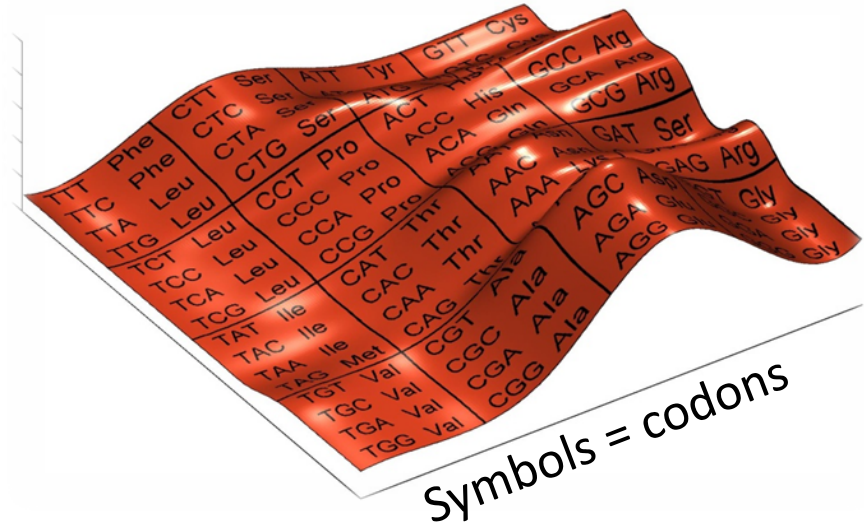


- Amino acid = backbone + specific side group.
- Diversity of amino acids allows proteins to perform a wide variety of functions efficiently.

The Genetic Code is highly ordered

TTT	Phe	CTT	Ser	ATT	Tyr	GTT	Cys
TTC	Phe	CTC	Ser	ATC	Tyr	GTC	Cys
TTA	Leu	CTA	Ser	ATA	TER	GTA	TER
TTG	Leu	CTG	Ser	ATG	TER	GTG	Trp
TCT	Leu	CCT	Pro	ACT	His	GCT	Arg
TCC	Leu	CCC	Pro	ACC	His	GCC	Arg
TCA	Leu	CCA	Pro	ACA	Gln	GCA	Arg
TCG	Leu	CCG	Pro	ACG	Gln	GCG	Arg
TAT	Ile	CAT	Thr	AAT	Asn	GAT	Ser
TAC	Ile	CAC	Thr	AAC	Asn	GAC	Ser
TAA	Ile	CAA	Thr	AAA	Lys	GAA	Arg
TAG	Met	CAG	Thr	AAG	Lys	GAG	Arg
TGT	Val	CGT	Ala	AGT	Asp	GGT	Gly
TGC	Val	CGC	Ala	AGC	Asp	GGC	Gly
TGA	Val	CGA	Ala	AGA	Glu	GGA	Gly
TGG	Val	CGG	Ala	AGG	Glu	GGG	Gly

meaning = polarity



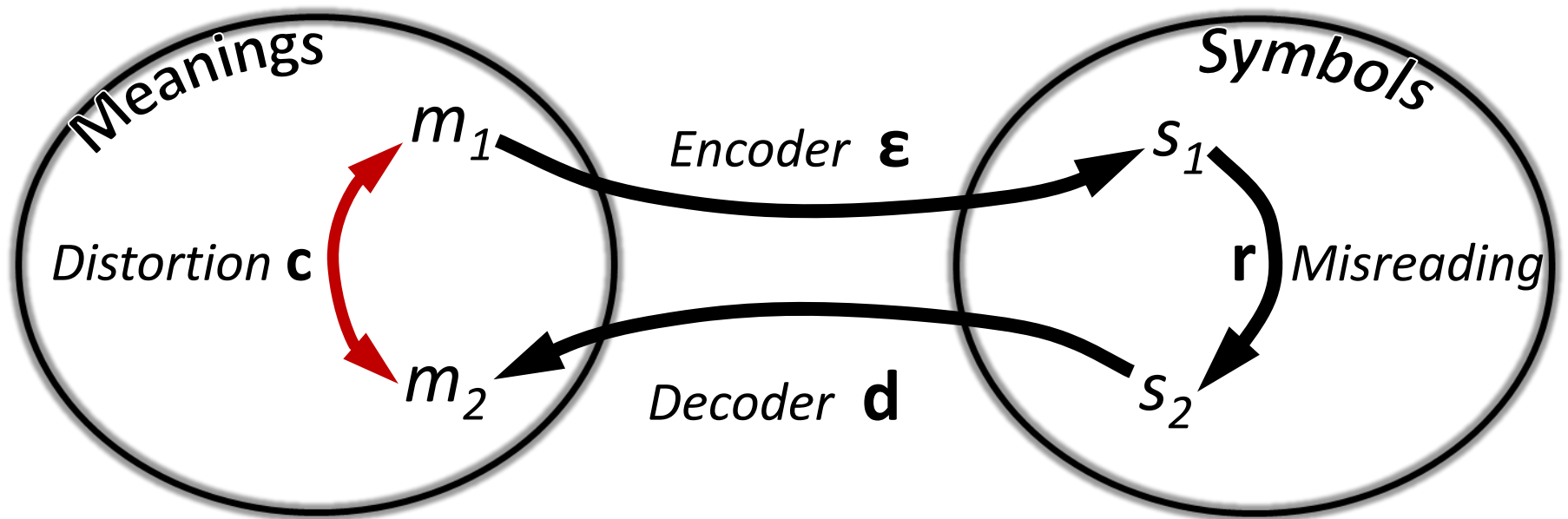
- Smooth.
- Degenerate (20 out of 64)
- Yet, diverse.
- Generic properties ?

Q: How do smooth codes emerge and evolve ?

A: Molecular codes are smooth

(1) to withstand noise (2) at a minimal cost.

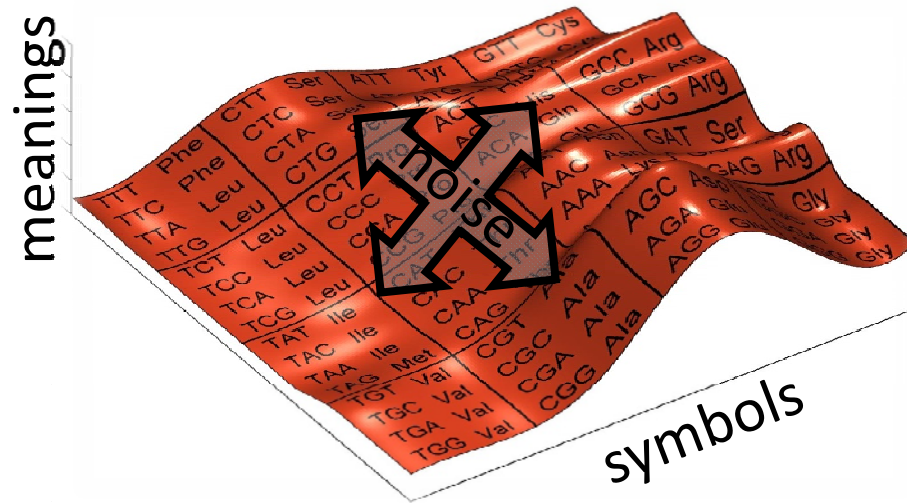
Molecular code is a channel with a quality measure



$$\text{Quality: } D = \langle c \rangle = \text{Tr}(\epsilon \cdot \mathbf{r} \cdot \mathbf{d} \cdot \mathbf{c})$$

- The quality **D** is the average distortion **$\mathbf{c}$** of a typical meaning.
- **ϵ** and **$\mathbf{d}$** determine the *quality* for a given misreading **$\mathbf{r}$** .
- **$\mathbf{r}$** defines the *topology* of symbol space.

Smooth codes minimize error-load

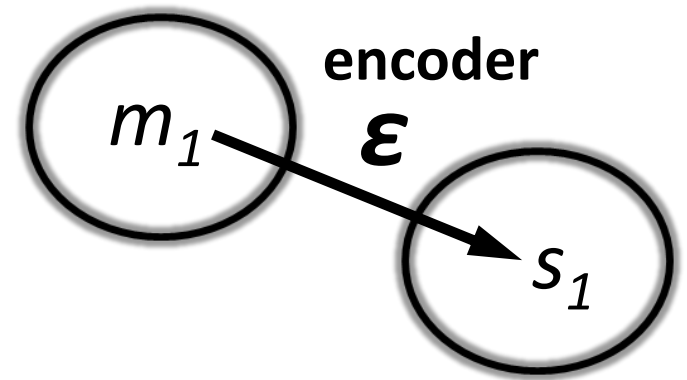


- Errors (noise) confuse similar, neighboring symbols.
- Smooth codes → neighboring symbols are also similar in meaning.
- → minimal impact of errors.
- Reading $\mathbf{r} \sim$ Laplacian operator in symbol space, Δ_s .
- Quality $\mathbf{D} \sim$ “elastic” energy of symbol space with meanings metric.

Molecular codes cost chemical specificity

- To encode/decode diverse meanings, molecular readers require **specificity** = high binding energies E_b .
- Cost $I \sim$ average binding energy $\langle E_b \rangle$.
- Binding probability $\sim$ Boltzmann: $\epsilon \sim \exp(E_b/T)$.

$$I = \langle \epsilon \ln \epsilon \rangle_{\text{encoder}} .$$



- Specificity cost is measured by mutual information I .

Code's fitness combines quality and cost

- *Quality = Error-load + Diversity*

$$D = \text{Tr}(\varepsilon \cdot \mathbf{r} \cdot \mathbf{d} \cdot \mathbf{c})$$
$$+$$

- *Cost = Chemical Specificity*

$$I = \langle \varepsilon \ln \varepsilon \rangle$$

$$\text{Fitness} = \text{Quality} + \text{Cost/Gain}$$

$$H = D + \kappa^{-1} I$$

- **Gain** κ increases with **complexity** of organism
and **richness** of environment.

Survival of the fittest code

- Population of “organisms” that compete and evolve according to code fitness, H .
- Population dynamics:

Max(quality)

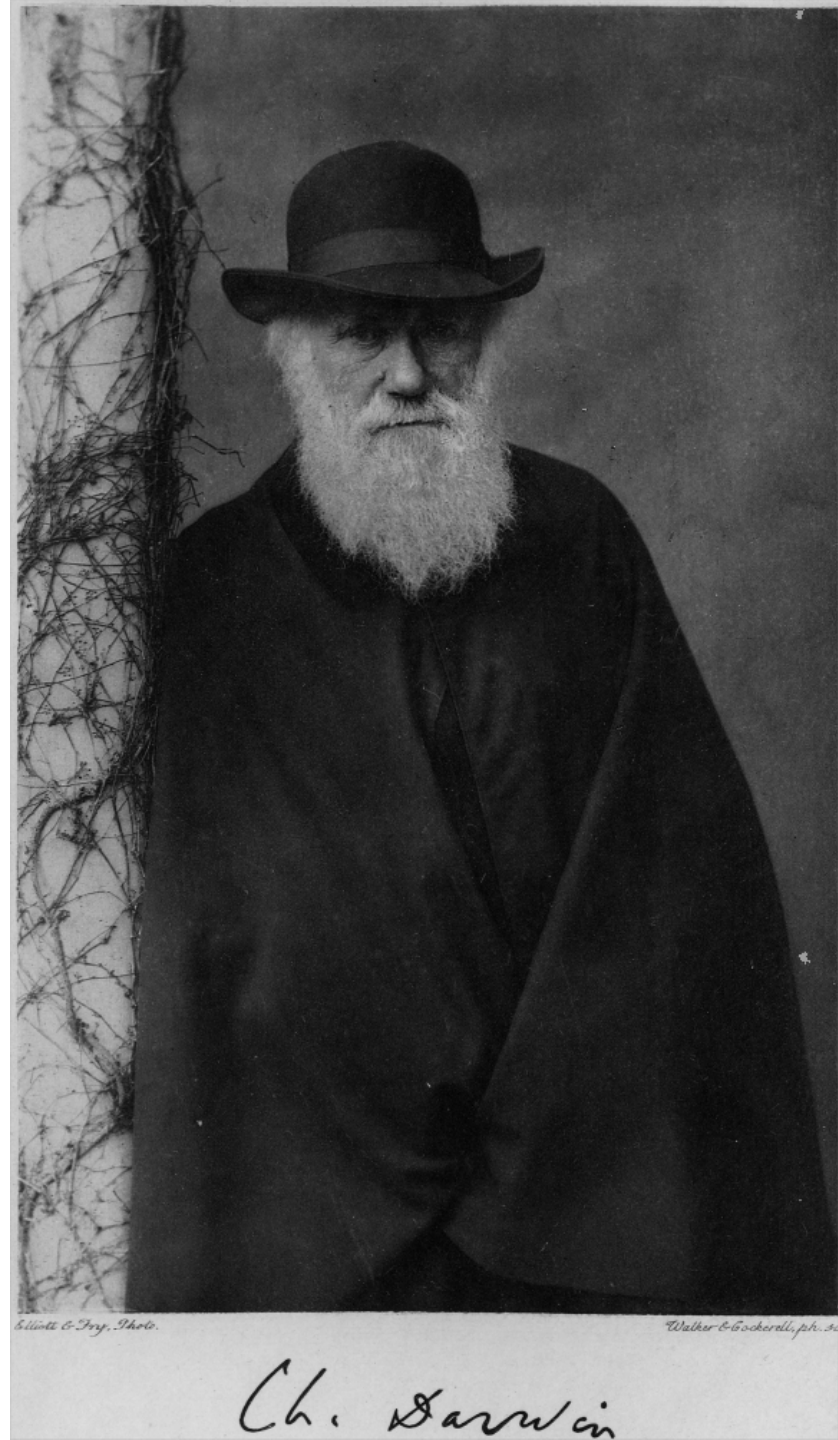
Min (cost)

+

Mutation

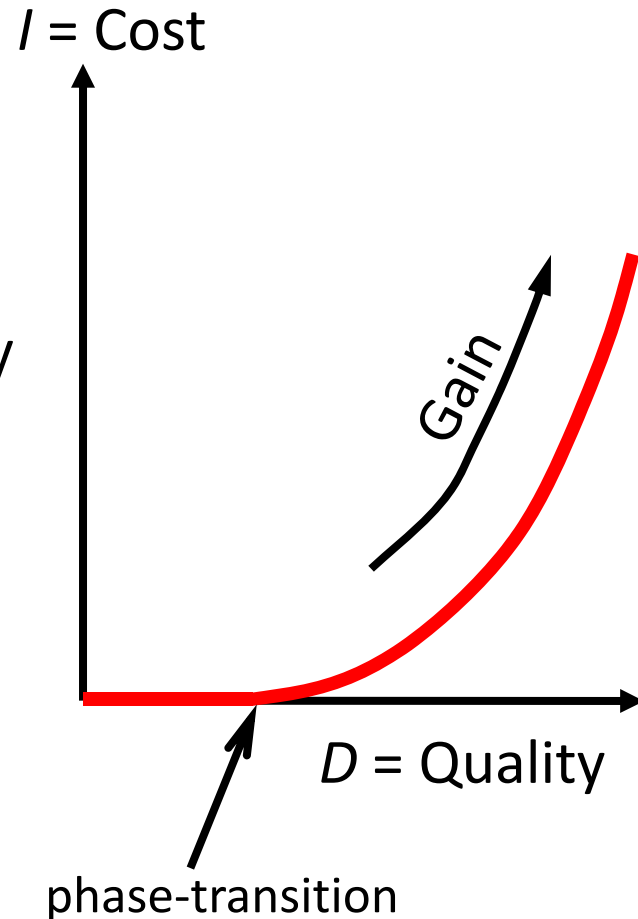
Selection

Random drift



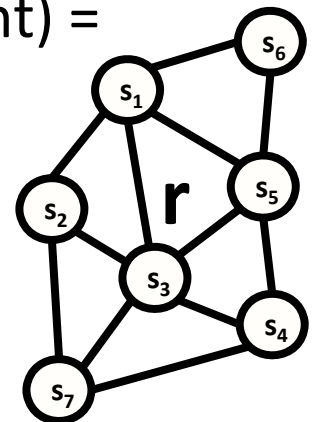
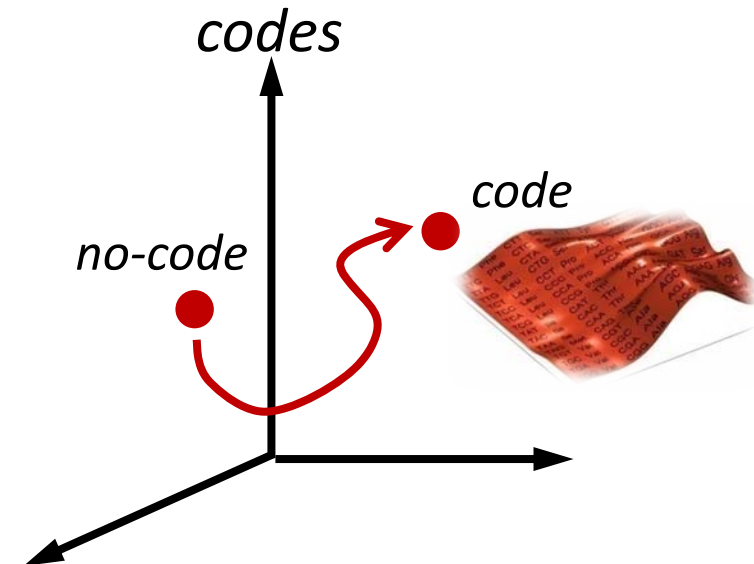
A code is born when gain increases

- Low gain: Cost too high $\rightarrow$
no specificity $\rightarrow$ no correlation $\rightarrow$ **no code**.
- **Code emerges** when channel starts to convey
info between symbols and meanings ($I \neq 0$).
- Instability of $H \rightarrow$
Continuous 2nd order phase transition.

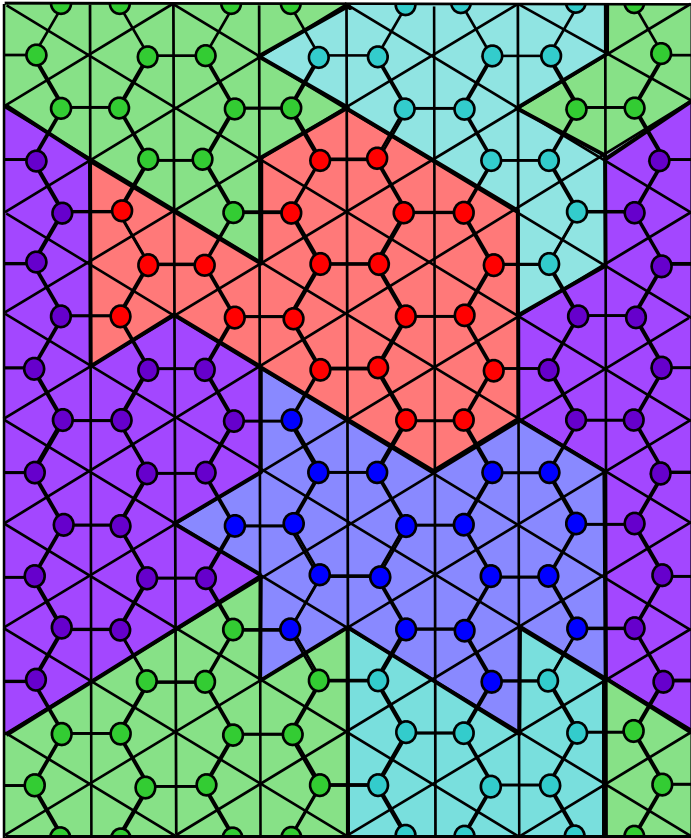


Codes appear as smooth modes in symbol space

- Instability of H ($\sim$ free energy)
→ phase transition
- Code = *Smoothest* non-uniform correlation pattern.
- Code is smooth 2nd mode of symbol Laplacian (Courant) = minimal “surface tension” of meaning islands.
- Misreading $\mathbf{r}$ is the graph-Laplacian $\mathbf{r} \sim \Delta_S$.



Optimal coding is a topological coloring problem

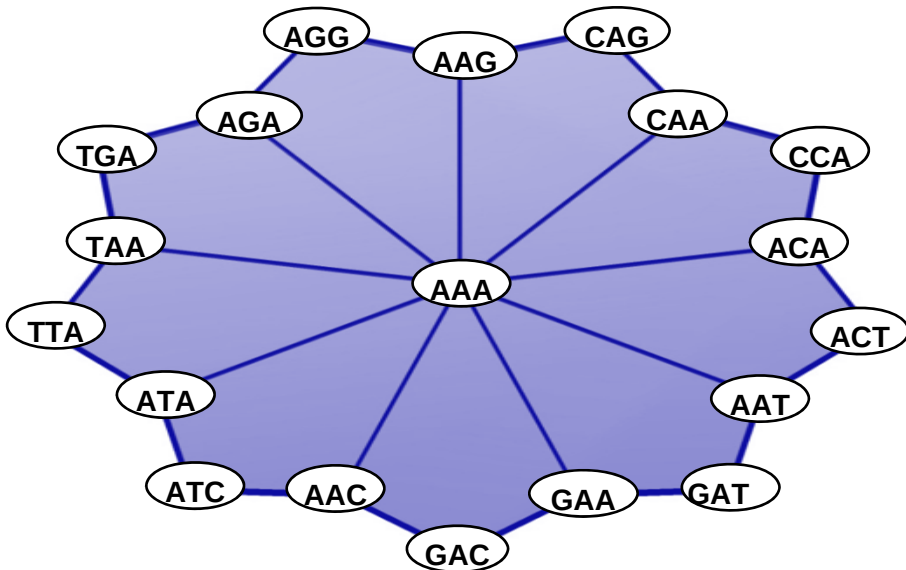
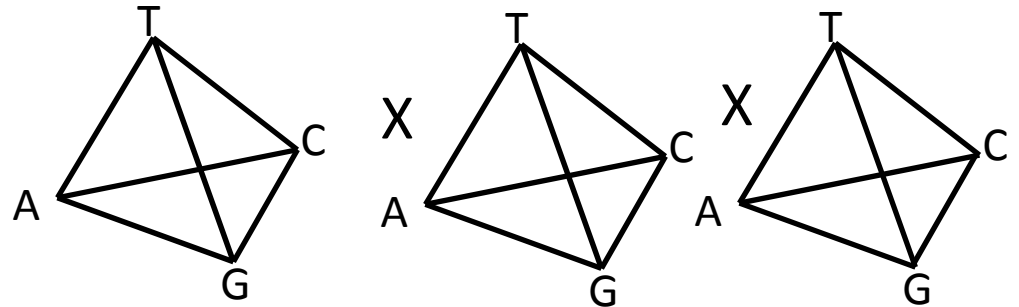


- Each color denotes a meaning (for example an amino-acid).
- Coloring partitions the symbol space.
- The code is optimal when every color or meaning has one compact contiguous island of words.
- Partition described by statistical mechanics of polymer networks.

The probable errors define the graph and the topology of the genetic code

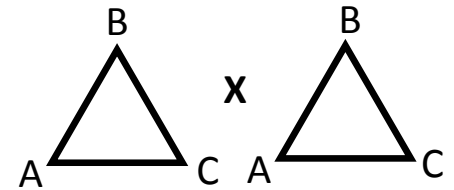
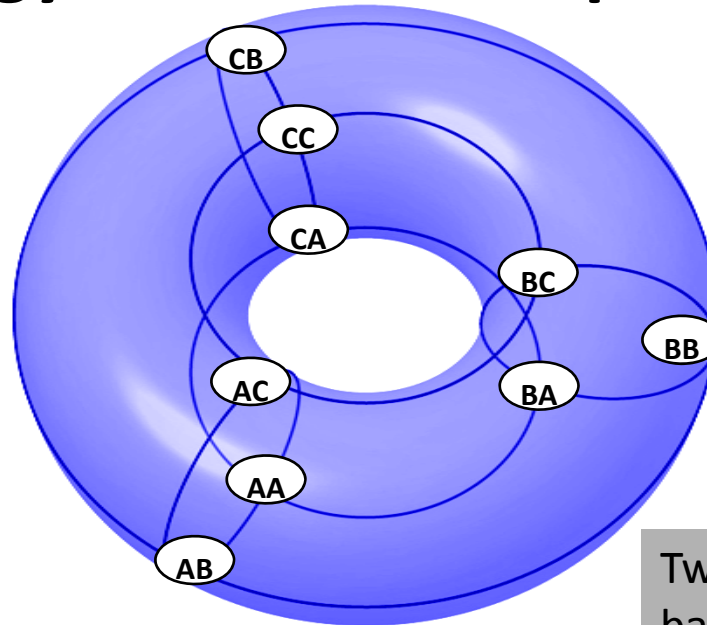
- Symbol (codon) Graph = codon vertices + one-letter difference edges (Hamming = 1)

$$K_4 \times K_4 \times K_4$$



Topology of a much simpler code

AA	BA	CA
AB	BB	CB
AC	BC	CC

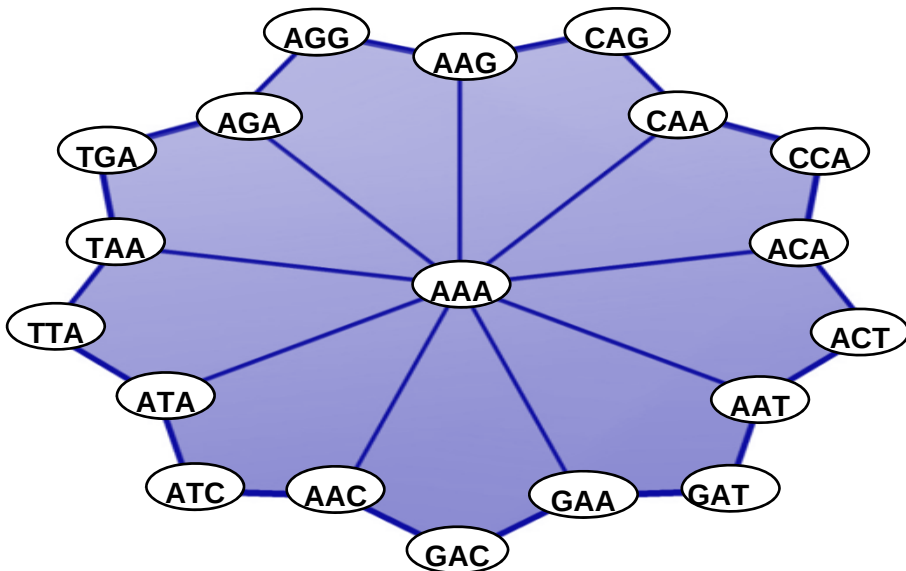
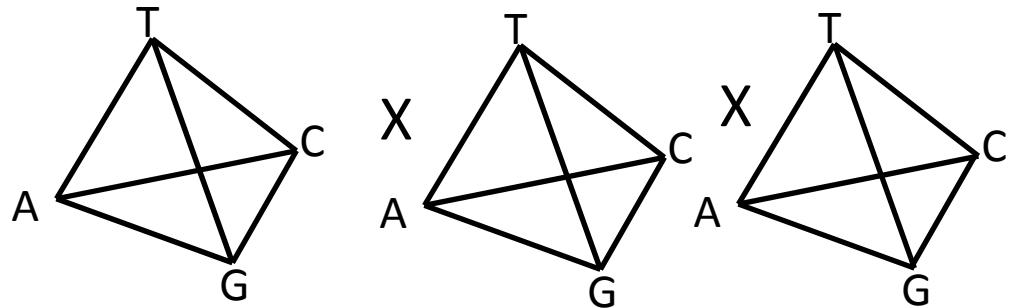


Two letter symbols with 3 bases is embedded on a torus.

- Euler's characteristic: $\chi = \text{Vertices} - \text{Edges} + \text{Faces}$.
- Genus (# holes): $\gamma = 1 - \chi/2$.
- Faces are quadrilaterals:
 $\text{Vertices} = \text{Faces} = 9$; $\text{Edges} = 18$.

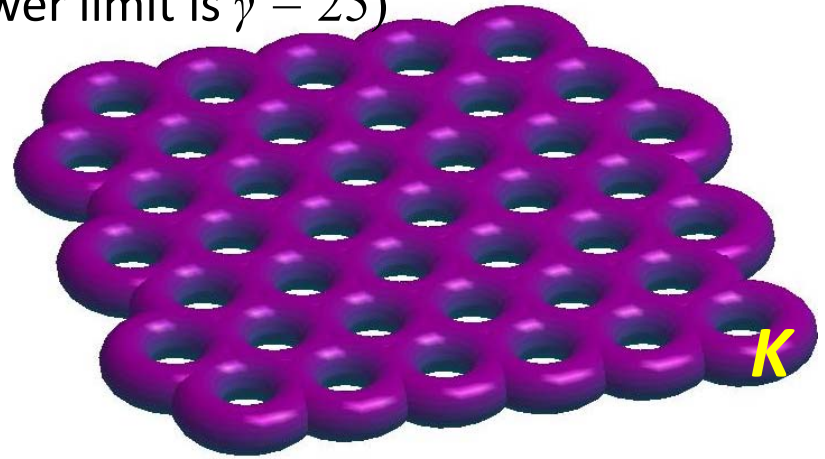
The surface of the code graph is holey

$$K_4 \times K_4 \times K_4$$



Holey graph: $\gamma = 41$

(lower limit is $\gamma = 25$)



Coloring number is the upper limit for the number of smooth islands

- What is the minimal number of colors required for a map so that no two adjacent countries have the same color?
- Coloring number is a topological invariant and a function of the genus, $chr(\gamma) = \left\lfloor \frac{1}{2} \left(7 + \sqrt{1 + 48\gamma} \right) \right\rfloor$.

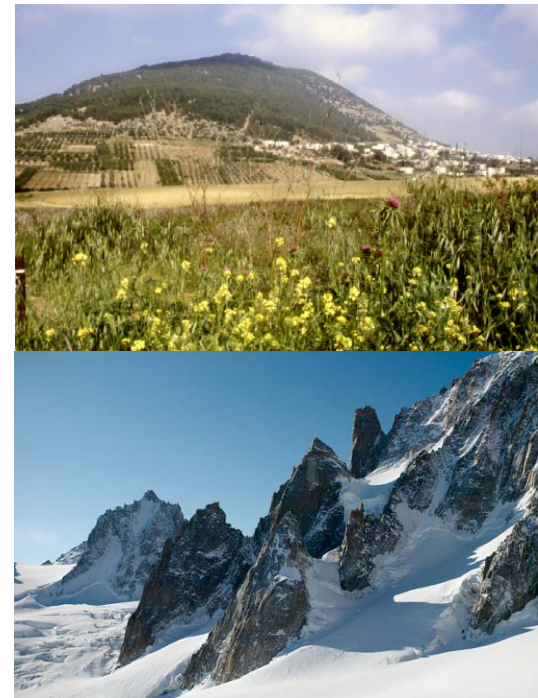
of meanings = $chr(\gamma)$

4	7	8	9	10	11	12	12	13	13
14	15	15	16	16	16	17	17	18	18
19	19	19	20	20	20	21	21	21	22
22	22	23	23	23	24	24	24	24	25
25	25	25	26	26	26	27	27	27	27

Topology determines the optimal coloring

- Each meaning has single compact domains with one maximum and one minimum (Courant).
- Compact organization reduces impact of errors.
- Embedding in R^{N-1} is **tight** or 'convex'
→ The code graph contains complete graph K_N

$$\# \text{ meanings} = N = \text{coloring}(\gamma)$$



Other molecular codes:

Transcription regulatory network:

- Controls gene expression via binding proteins to DNA.
- Mapping between proteins and DNA is.
- Number of proteins is limited by the coloring number.

Logic design of operons:

- Logic gates made of binding proteins are smooth.

Probable recognition errors define the binding sequence space

- Coloring number estimate:

$$v = 4^L \quad (L=6)$$

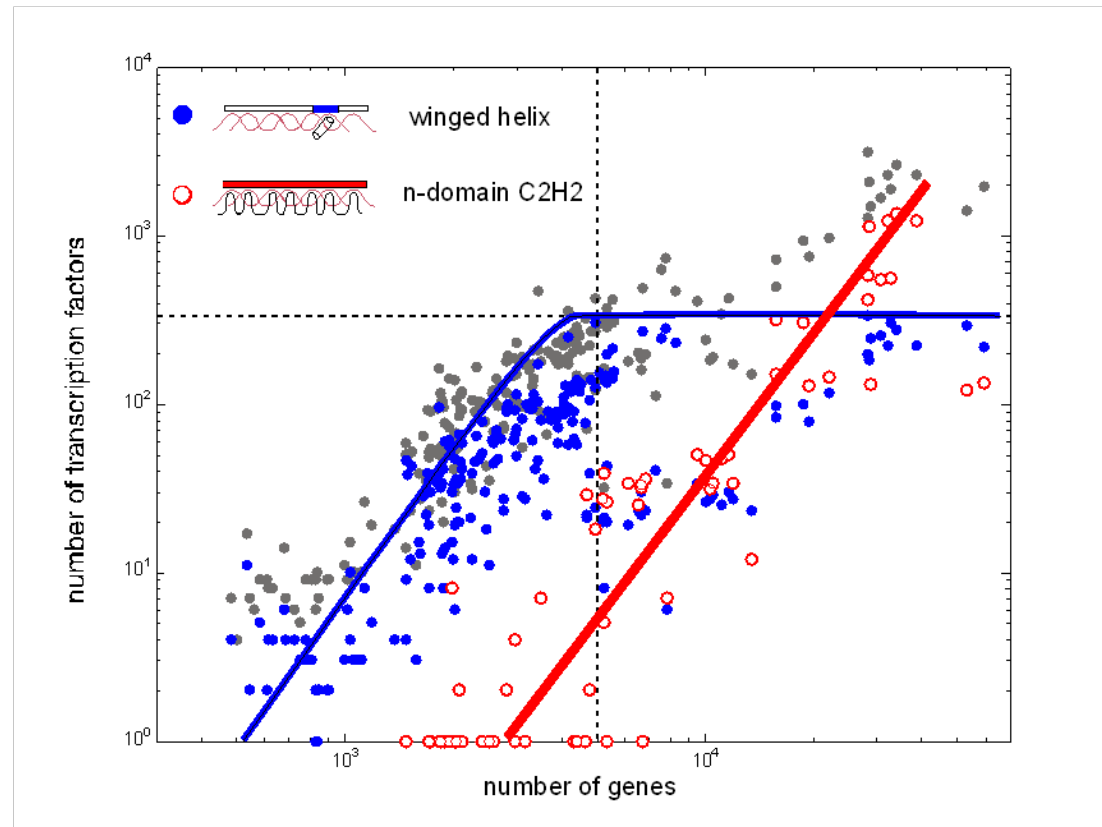
$$e \sim 4^L(3/2)L$$

$$f \sim 4^L(3/4)L$$

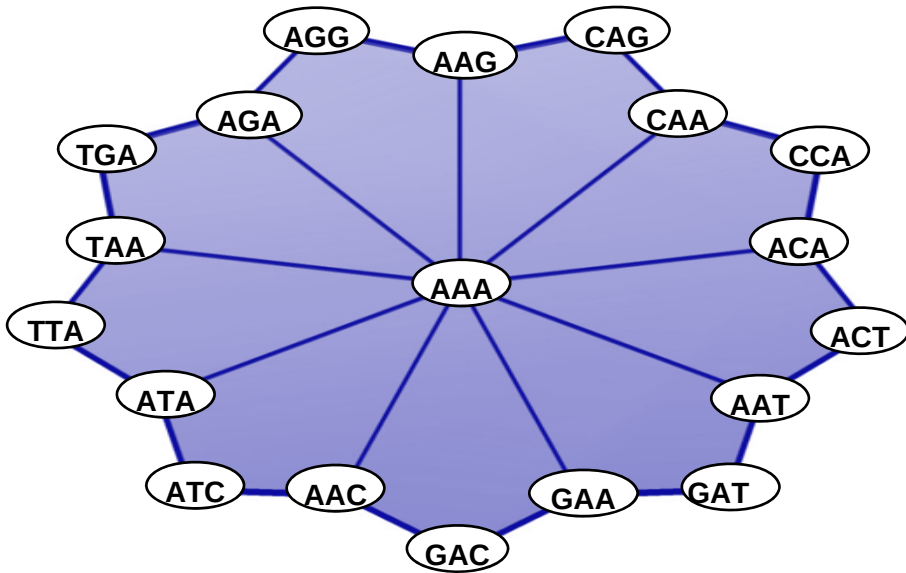
$$\rightarrow \gamma \sim 4^L(3/8)L$$

- The coloring #

$$\text{chr}(\gamma) \sim 300$$



Optimal coding is a topological coloring problem – optimizing number of meanings



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

- Topology of error-Laplacian $\mathbf{r}$ governs coding transition.
- Smoothness limits number of meanings due to tightness of map.
- The limit is the coloring number, determined by topology (γ).
- Genetic code $\gamma = 25-41 \rightarrow$ coloring number = 20-25 amino-acids.

Population dynamics: mutations, genetic drift

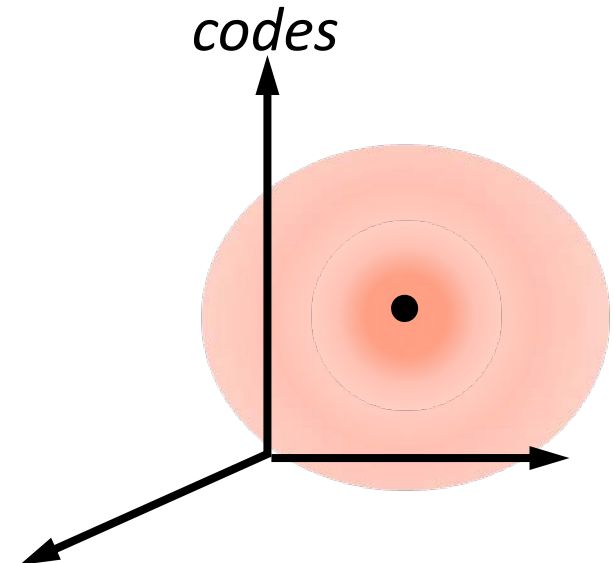
- Mutations smear the population in code-space.
- Reaction-diffusion dynamics reach steady-state

$$\frac{\partial \Psi}{\partial t} = -H\Psi + \mu \nabla_{code}^2 \Psi$$

(μ - mutation rate, ψ - population density)

$$\Psi \sim \exp\left(-\mu^{-1/2} \times \varepsilon \times \varepsilon\right)$$

- Other effects : Genetic drift = reproduction fluctuations.



Summary

- Molecular codes = *maps* or information channels with fitness.
- *Fitness* = *Quality* + *Cost*.
- *Smooth* codes emerge at *phase transitions*.
- *Topology* of errors governs emergent code.

Population dynamics: genetic drift

- Genetic drift = reproduction fluctuations = Noise.
- The population migrates between many possible optima.
- At steady state $P(H) \sim \exp(-H/T)$ [Sella and Hirsh]
with evolutionary temperature $\sim 1/(\text{population size})$.
- Effective free energy (Potts-like, or polymer net)

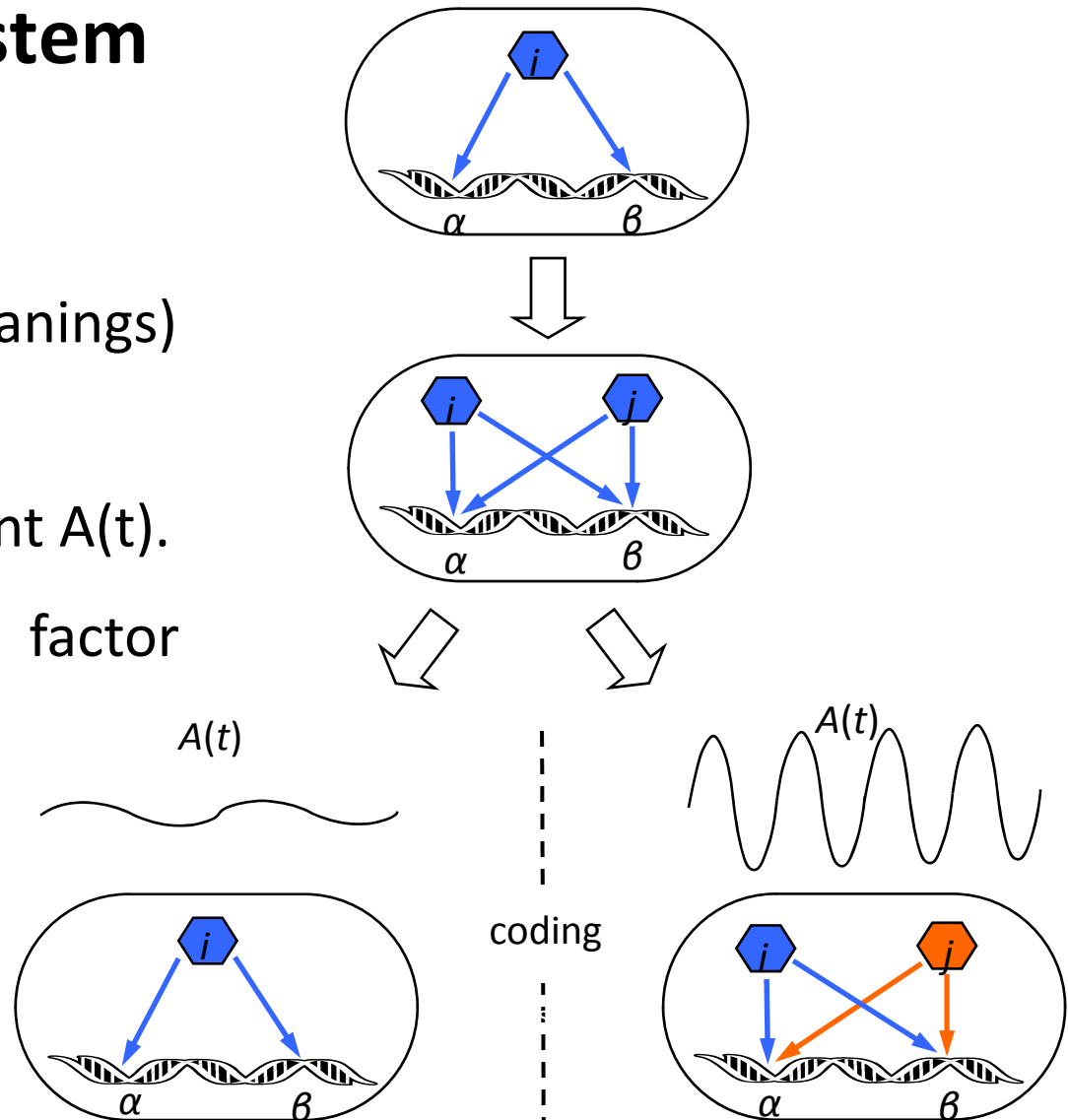
$$F = H(\varepsilon_{\alpha i}) + T \sum_{\alpha, i} \varepsilon_{\alpha i} \ln \varepsilon_{\alpha i}$$

- Shifting the critical gain

$$1 / \kappa_c + 1 / N = \lambda_r^2 \times \lambda_c$$

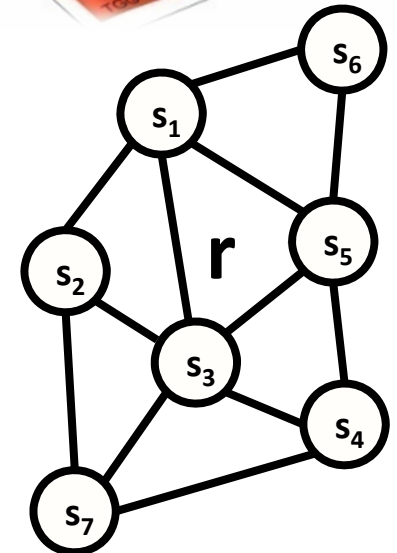
A sketch for an “experiment”: 2X2 coding system

- 2 binding sites (symbols).
- 2 transcription factors (meanings) after duplication.
- Control gain by environment $A(t)$.
- Coding transition when 2nd factor becomes advantageous.



Emergent code is a smooth mode of the error-Laplacian on symbol graph

- Every mode corresponds to a meaning
→ *number of modes = number of meanings.*
- Misreading $\mathbf{r}$ is the graph-Laplacian $\mathbf{r} \sim \Delta_s$.
- *Courant's theorem for Δ_s :*
single maximum for each *mode*
→ single contiguous domain for each meaning.
→ *Smoothness*



** Statistical mechanics of code evolution

- Fitness $H = D + I/\kappa$

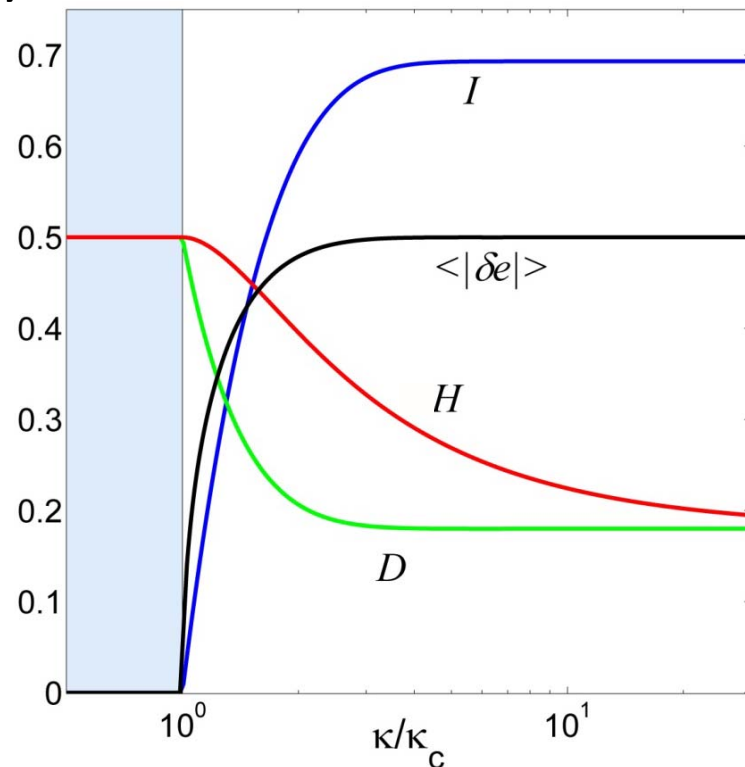
Quality + Cost/Gain $\sim$ Free energy

- Gain $\kappa \sim$ inverse temperature.
- Fittest code takes over.
- Given $\mathbf{r}, \mathbf{c}, \kappa$:

$$\min_{\{e,d\}} H \rightarrow \text{fittest code } (e^*, d^*).$$

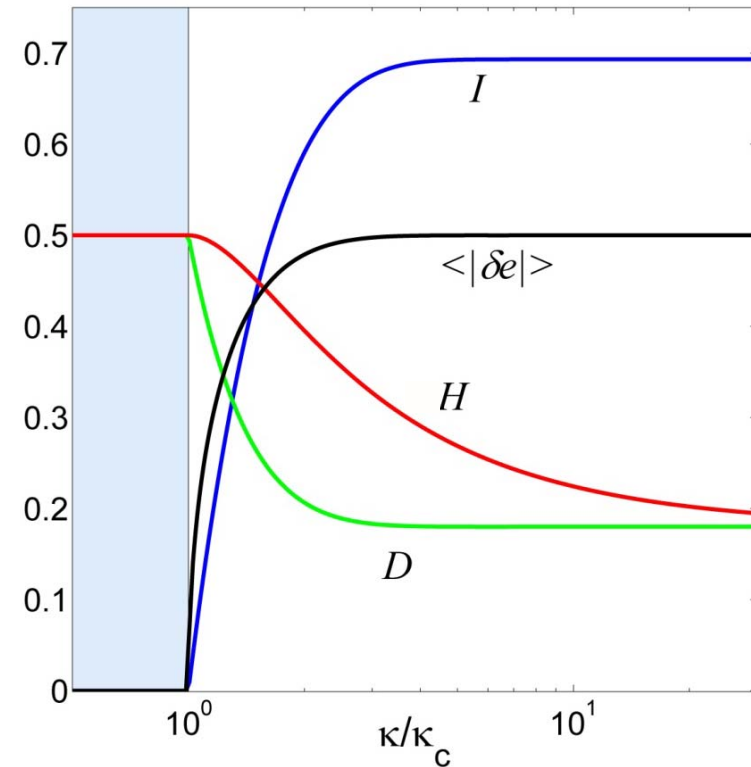
- Order parameter

δe = deviation from randomness.



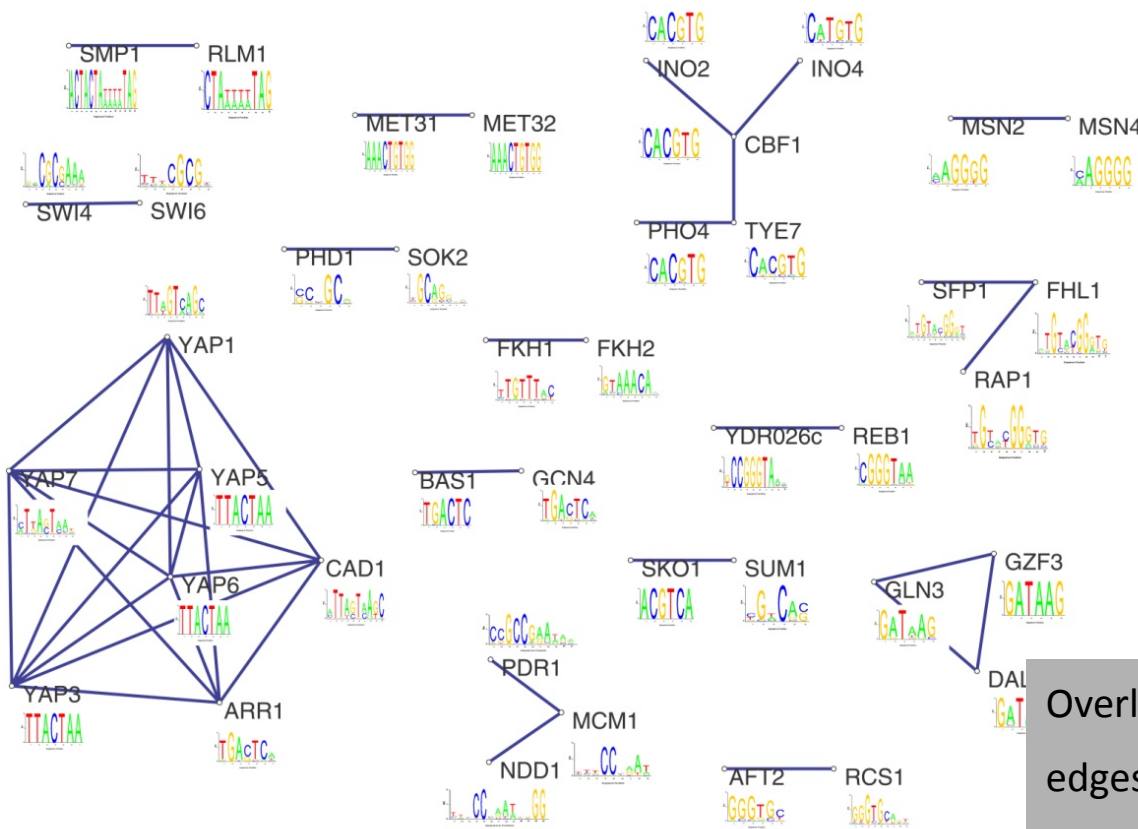
** Code emerges at a 2nd order transition

- Transition at critical gain κ_c .
- Critical “temperature” depends on **r** and **c**: $1/\kappa_c \sim \lambda_r^2 \times \lambda_c$.
- Code is the smooth mode e_{ms} of H that corresponds to 2nd e.v. of Δ .
- Three pathways to transition:
 - increase gain.
 - increase accuracy.
 - increase diversity.



The transcription network is smooth

- Transcription factors that bind to similar DNA sequences tend to have similar meanings



- Meaning is measured by the GO annotation or co-regulation

Overlapping TFs in Yeast. Vertices are TFs. edges connect TFs with overlapping 'spheres'.