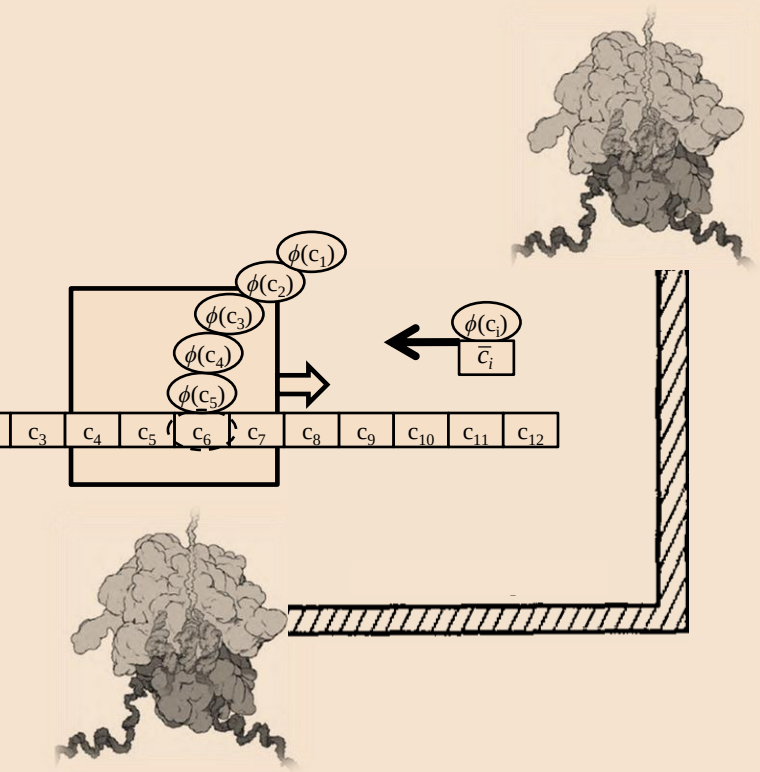
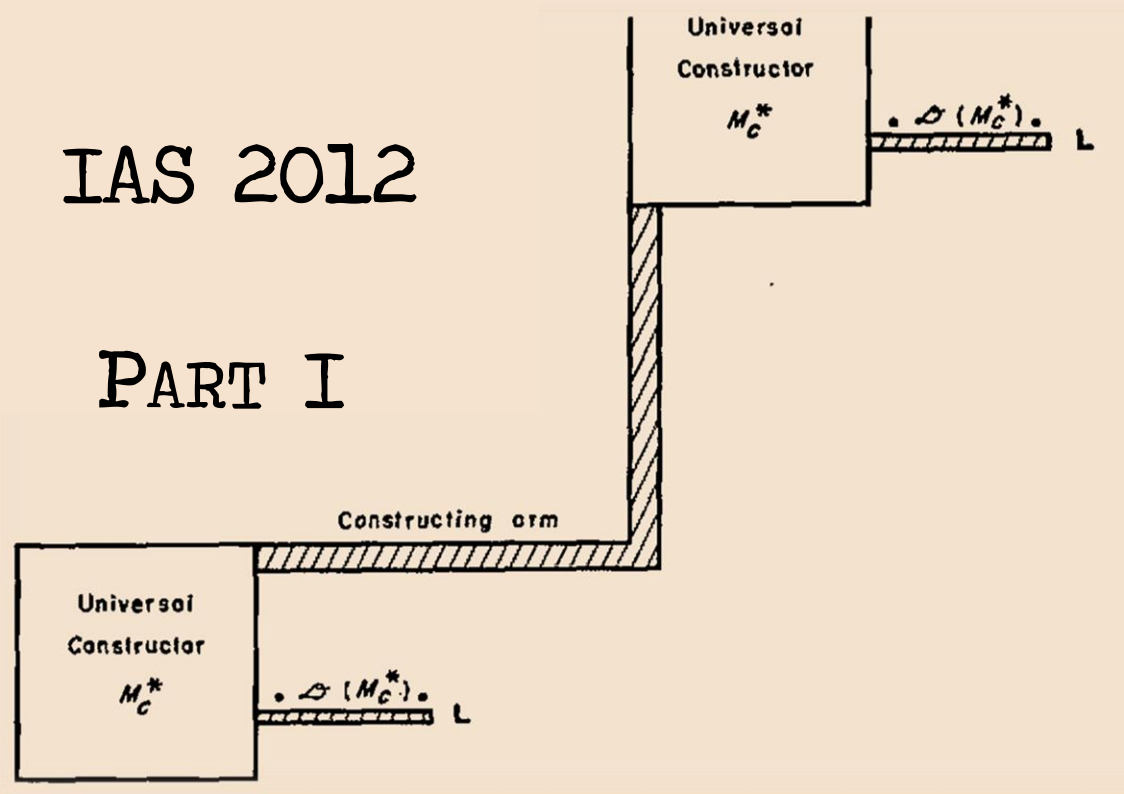
Coding a Life Full of Errors

PITP

IAS 2012

Part I

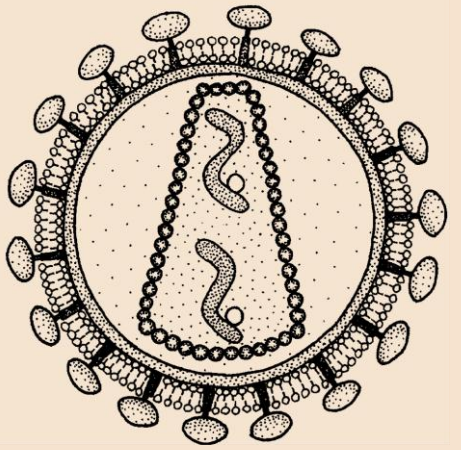


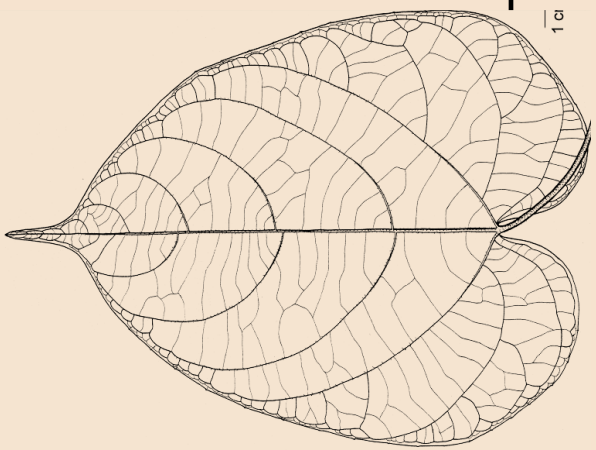


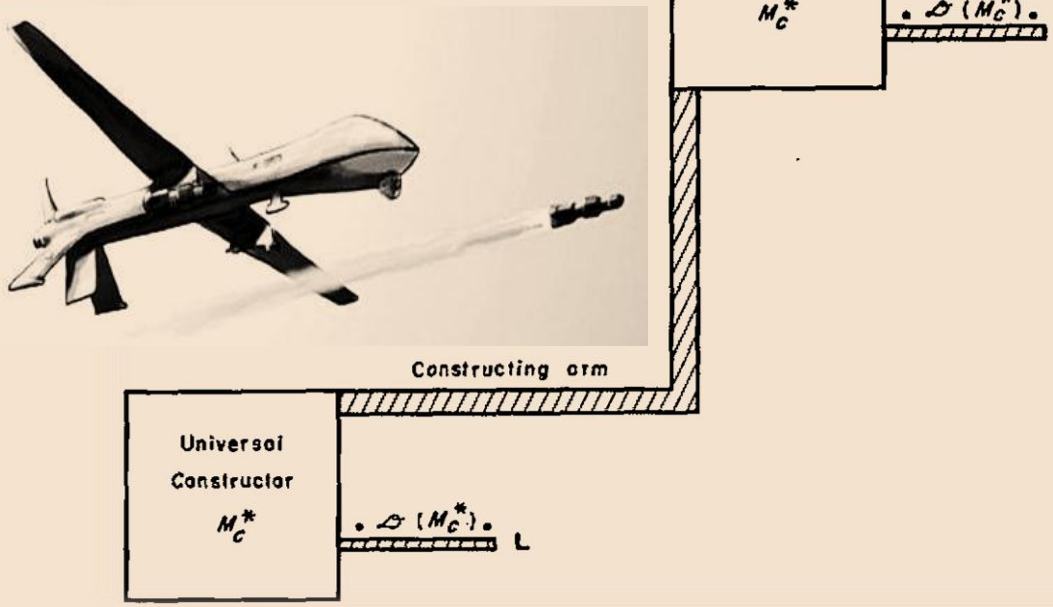
What is Life? (biological and artificial)

- Self-replication.
- Emergence.
- Evolution.
- Non-equilibrium.
- **Information.**
- **Geometry.**
- **Stochasticity**.

- Viruses (bio and computer).
- Growth and form.
- Natural algorithms .
- Learning & robots.
- **Codes and Errors.**







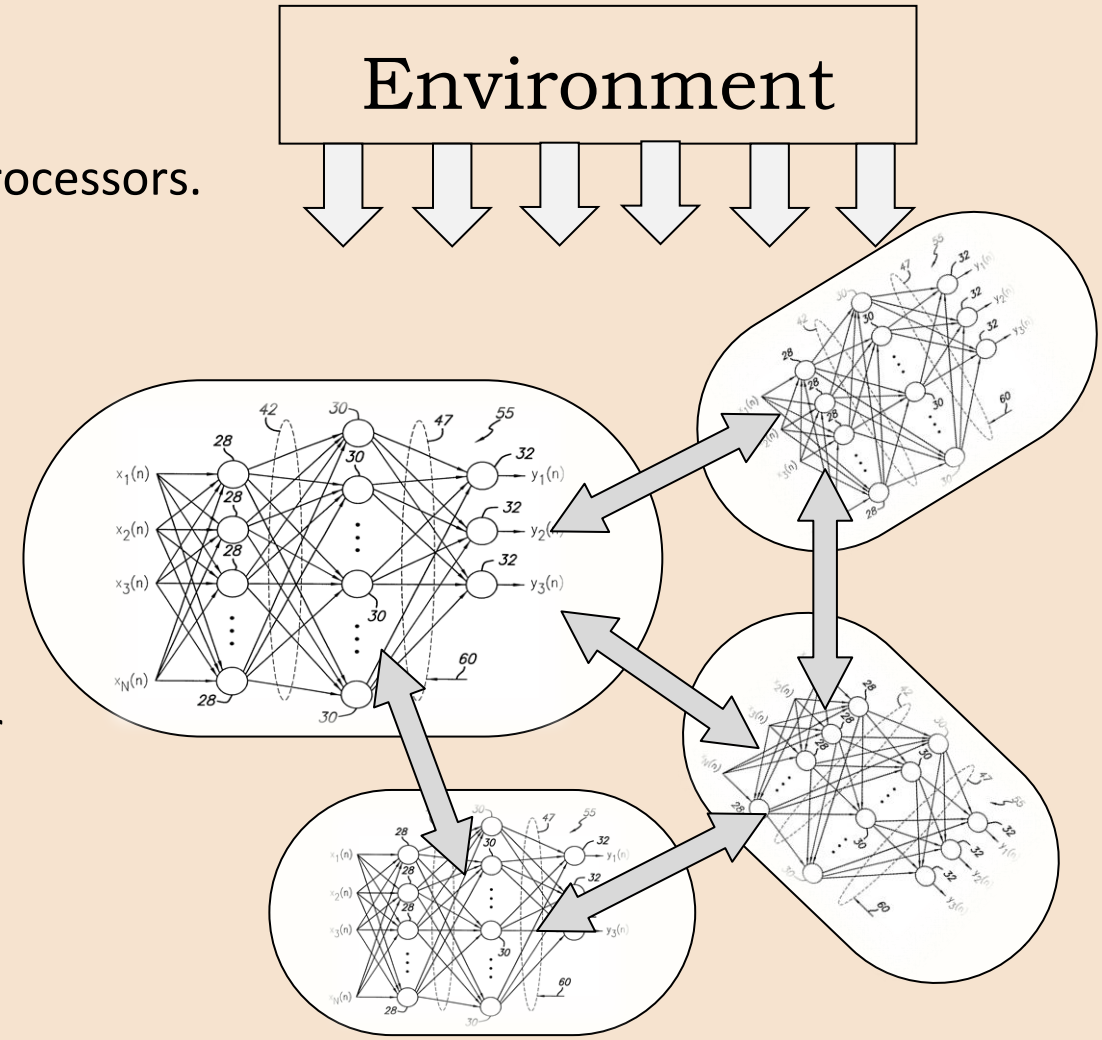
Living information is carried (mostly) by molecules

"Living systems"

I. Self-replicating information processors.

II. Evolve collectively.

III. Made of molecules.



- Generic properties of molecular codes subject to evolution?
- Information theory approach?

Challenges of molecular codes: rate and distortion

Distortion

- Noise, crowded milieu.
- Competing lookalikes.
- Weak recognition interactions ~ $k_B T$.
- Need diverse meanings.

"Synthesis of reliable organisms from unreliable components"
(von Neumann, *Automata Studies* 1956)

Rate

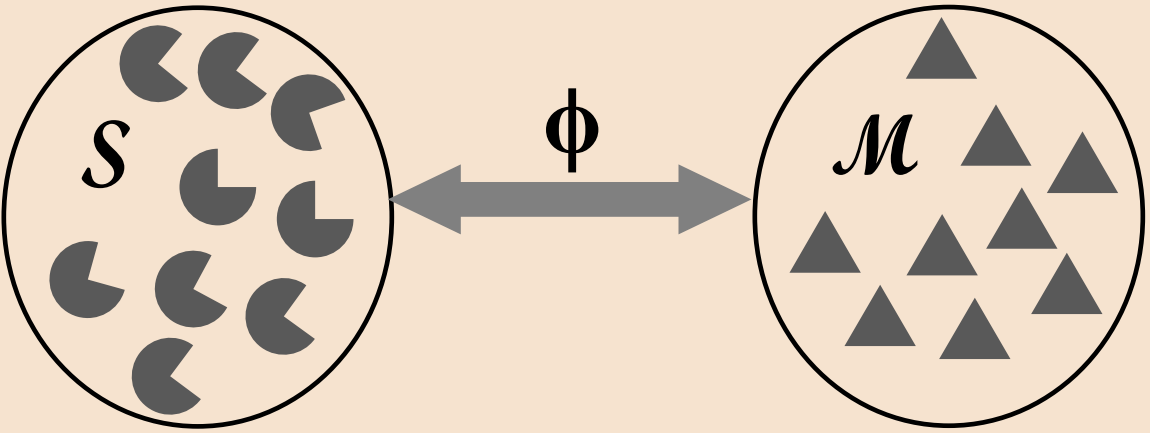
- How to construct the low-rate molecular codes at minimal cost of resources?

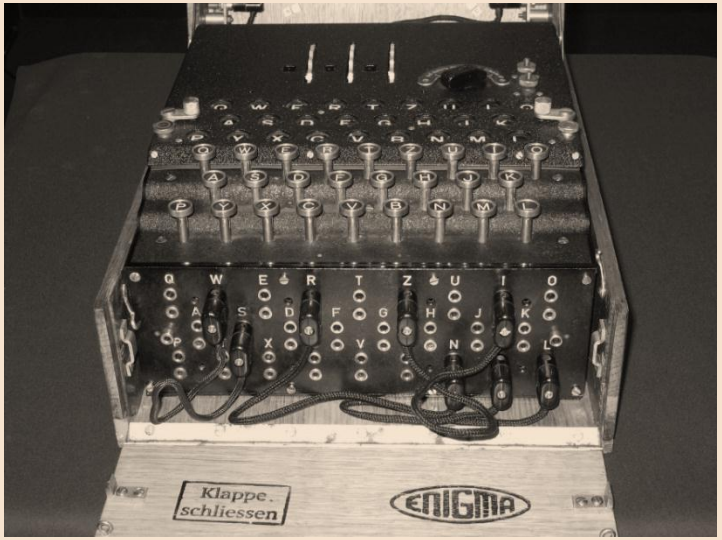
Rate-distortion theory (Shannon 1956)

Inside E. Coli, D. Goodsell

Codes are mappings, channels, representations, models…

- Code ϕ is a mapping between spaces, $\phi: S \rightarrow \mathcal{M}$.



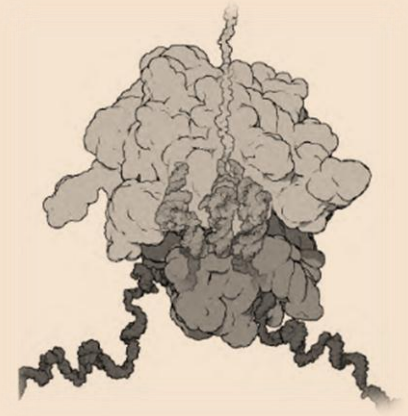


- Molecular codes map\translate between molecular spaces\languages.
- Molecular spaces have inherent geometry\topology.
- Coding machinery affects organism's fitness.

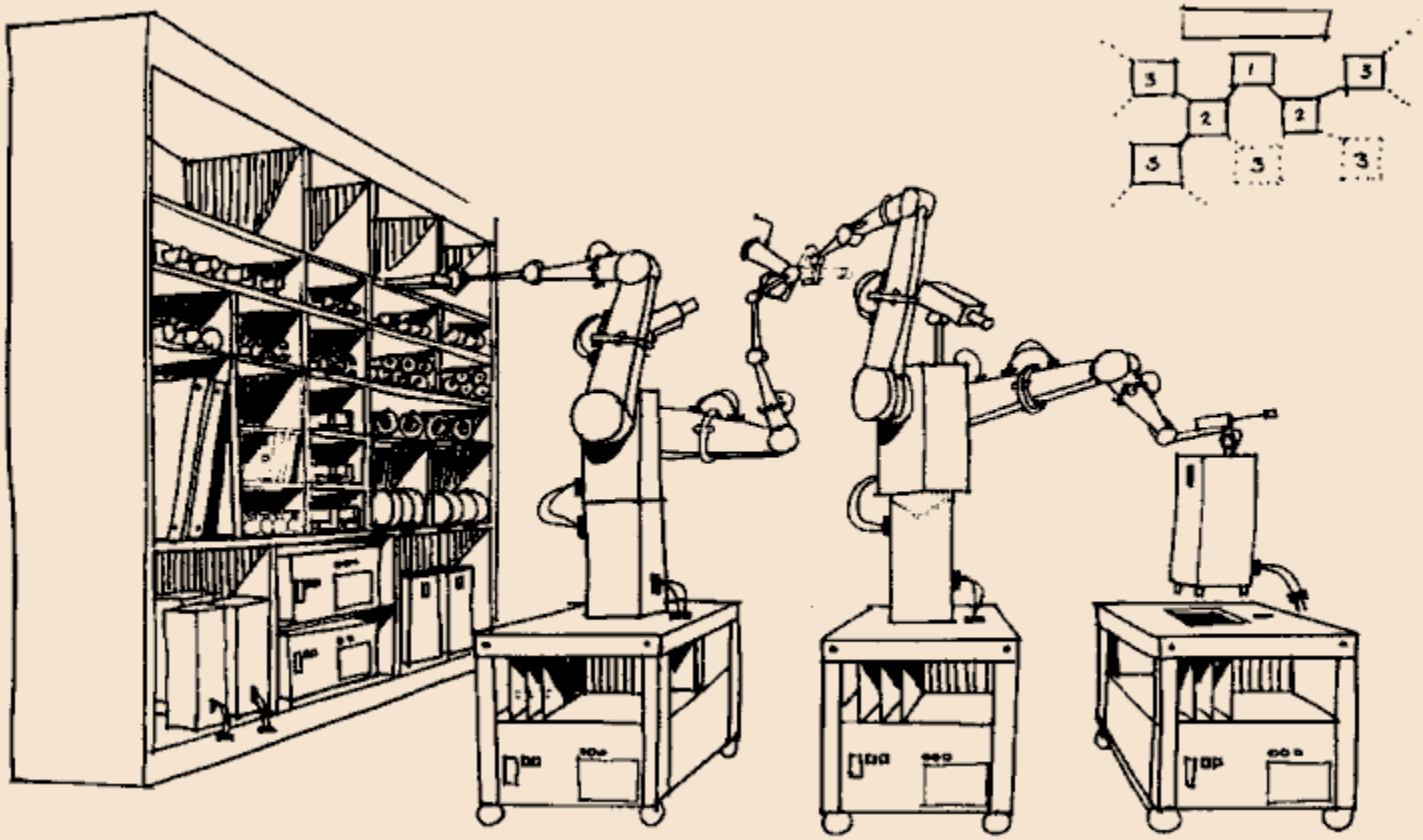
Outline: molecular codes and errors

- Living and artificial self-replication.
- The main molecular codes of life (central dogma).
- The translation machinery:
  - The genetic code, ϕ: codons → amino-acids.
  - The ribosome and the problem of molecular recognition.

- Basic coding theory: geometrical aspects.
  - How codes cope with errors.
- Emergence and evolution of codes: rate-distortion.
- Accuracy vs. rate: proofreading schemes.



Coding and the problem of self-replication



Proposed demonstration of simple robot self-replication,

from *advanced automation for space missions*, NASA conference 1980.

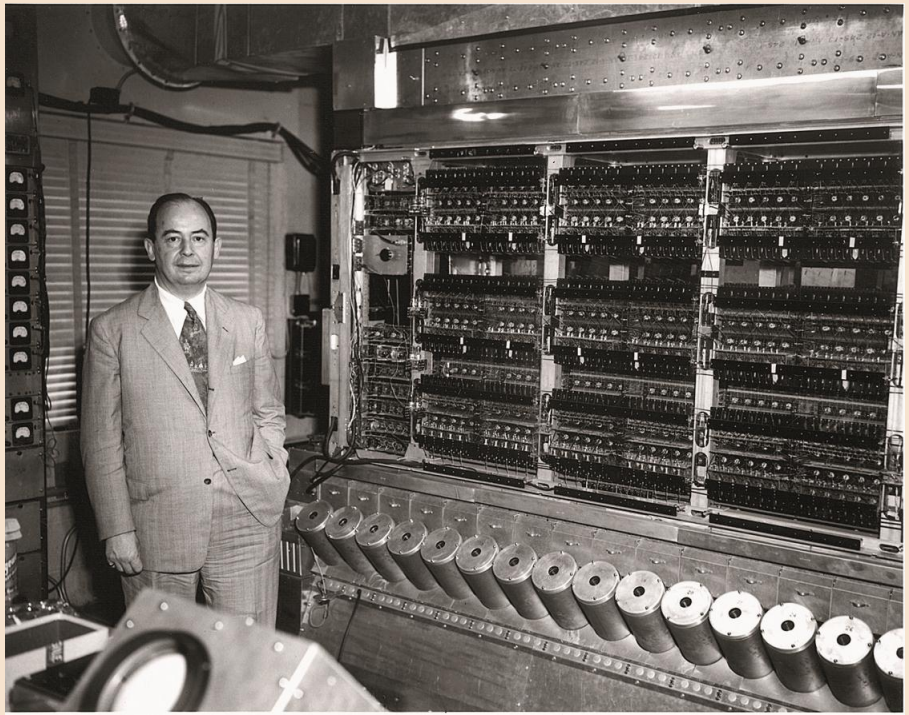
Self-replication and accuracy in computers

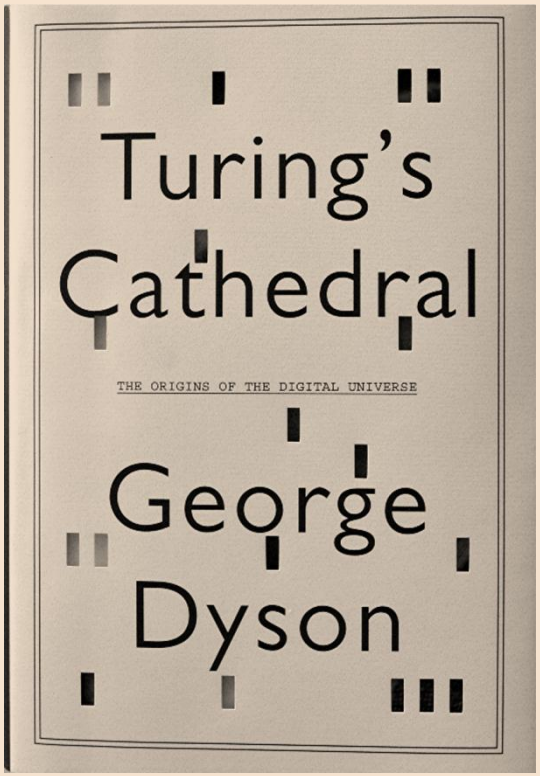
Theory of
Self-Reproducing Automata

JOHN VON NEUMANN
edited and completed by Arthur W. Burks

Lectures on

PROBABILISTIC LOGICS AND THE SYNTHESIS OF RELIABLE ORGANISMS FROM UNRELIABLE COMPONENTS





Theory of (1966)

Self-Reproducing Automata

Von Neumann's universal constructor

Self-reproducing machine: constructor + tape (1948/9).

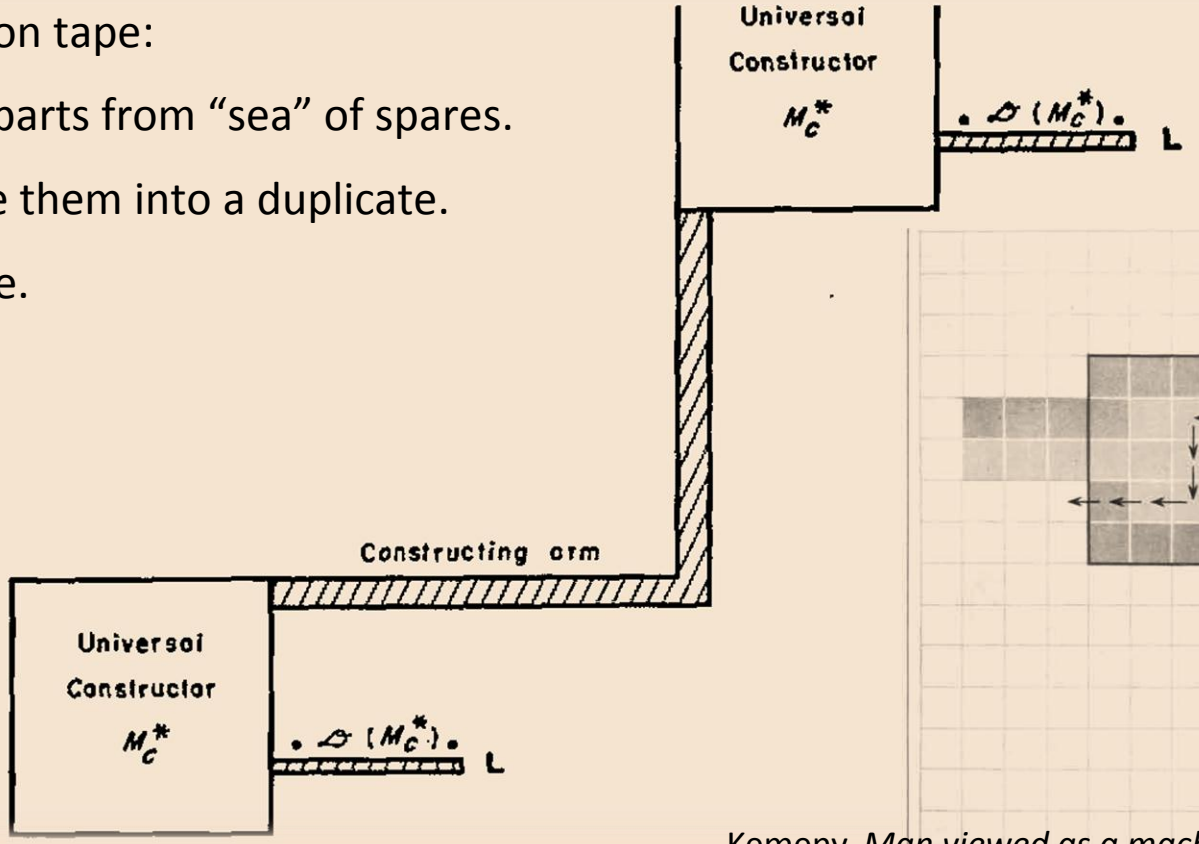
- Program on tape:

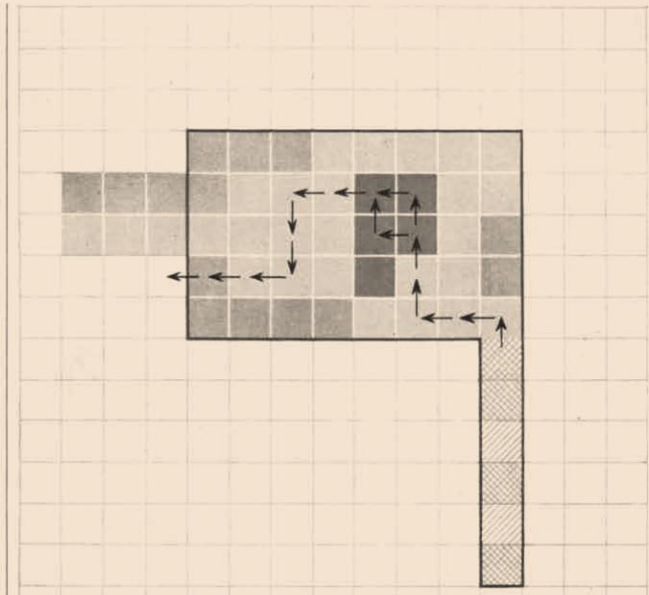
(i) retrieve parts from "sea" of spares.

(ii) assemble them into a duplicate.

(iii) copy tape.

.





Kemeny, *Man viewed as a machine , Sci Am* (1955)

Von Neumann's design allows open-ended evolution

Motivated by biological self-replication:

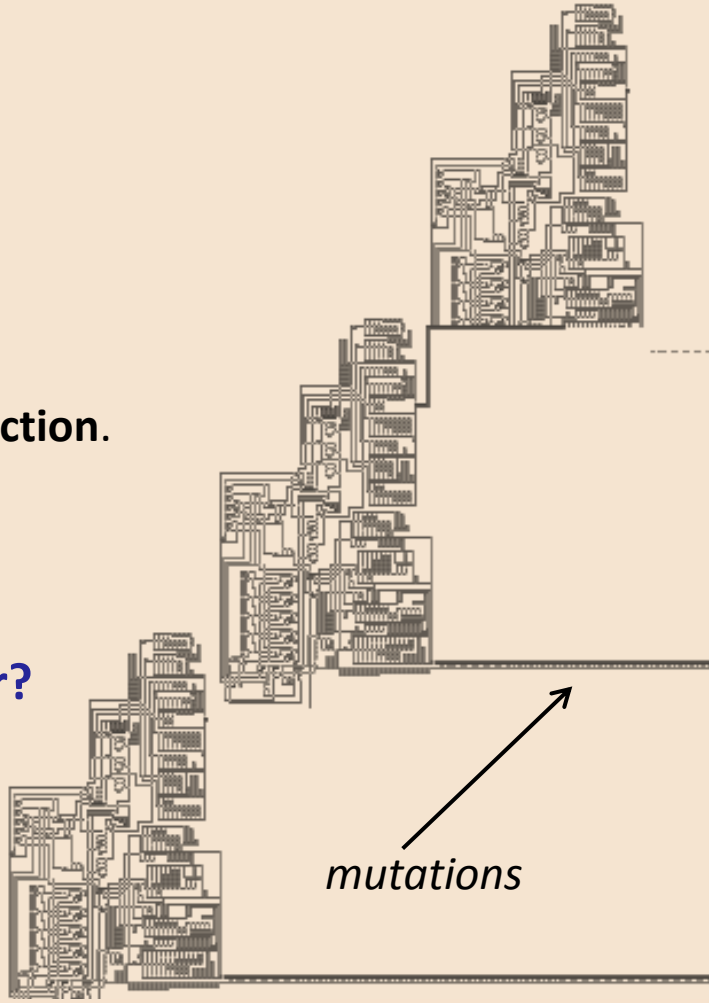
- Construction *universality.*
- *Evolvability*.

Key insight **(**before DNA) **separation of information and function**.

- Tape is read twice: for construction and when copied.

- **How to design fast/accurate/compact constructor?**

- **Requires efficient and accurate coding…**

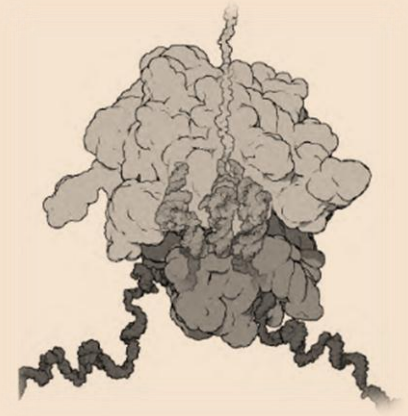


Implementation by Nobili & Pesavento (1995)

Outline: molecular codes and errors

- Living and artificial self-replication.
- **The main molecular codes of life** (central dogma).
- The translation machinery:
  - The genetic code, ϕ: codons → amino-acids.
  - The ribosome and the problem of molecular recognition.

- Basic coding theory: geometrical aspects.
  - How codes cope with errors.
- Emergence and evolution of codes.
- Accuracy vs. rate: proofreading schemes.

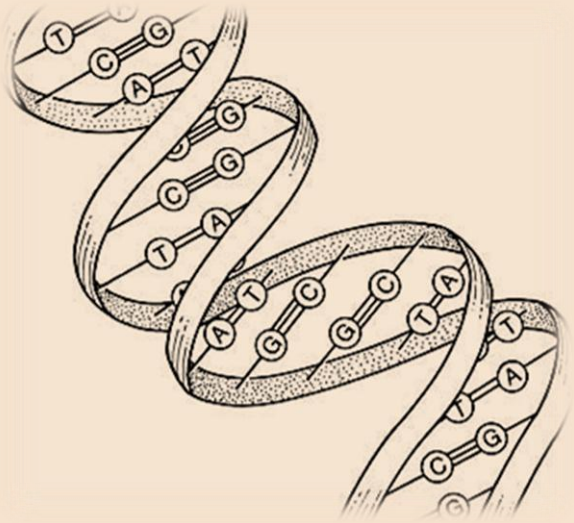


Dual spaces of DNA and proteins

DNA

- Building blocks :

  4 nucleic bases = {A, T, G, C}.

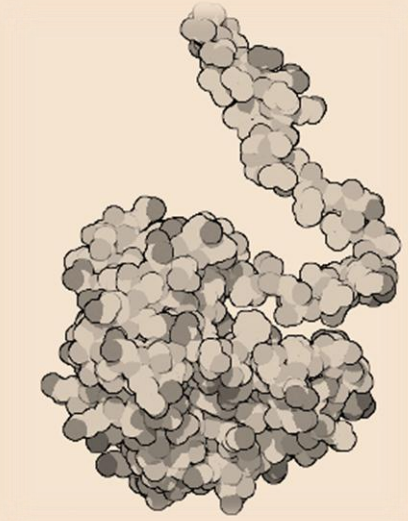


- Polymer: DNA double-helix.
- Inert information storage (“tape”)

protein

- Building blocks:

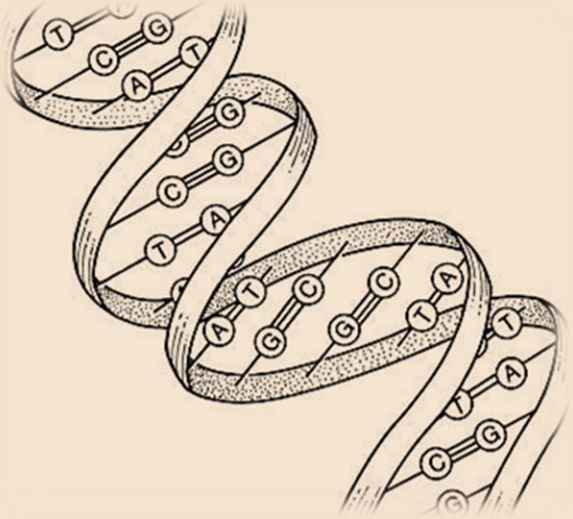
  20 amino acids.



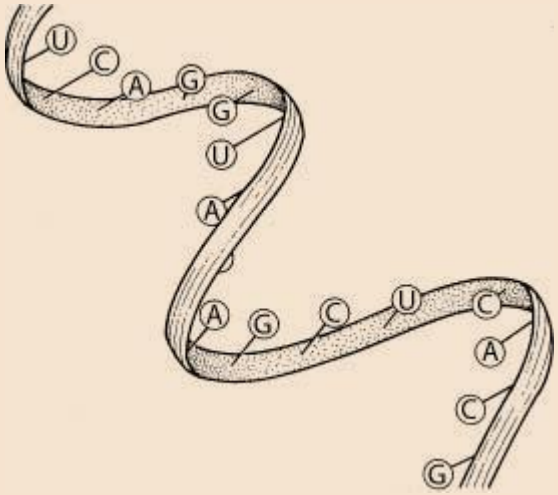
- Polymer = protein.
- Functional molecules (“constructor”)

RNA intermediates can be both tapes and machines

DNA



RNA



protein

- Primordial “RNA world” :

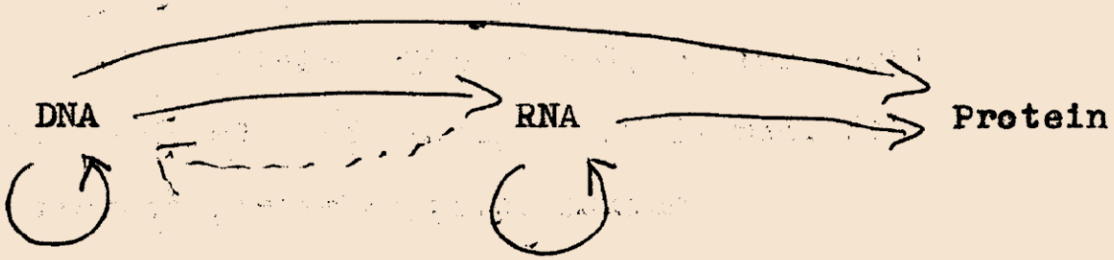
RNA molecules are both information carriers (DNA) and executers (proteins).

The Central Dogma of molecular biology

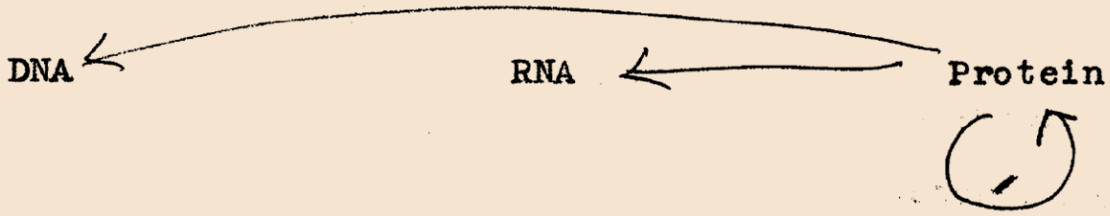
Francis Crick: Ideas on Protein Synthesis (Oct. 1956)

The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have



but never



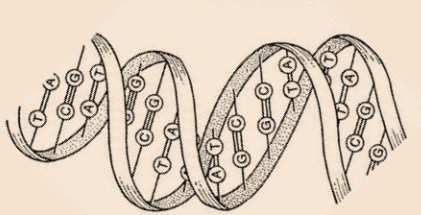
where the arrows show the transfer of information.

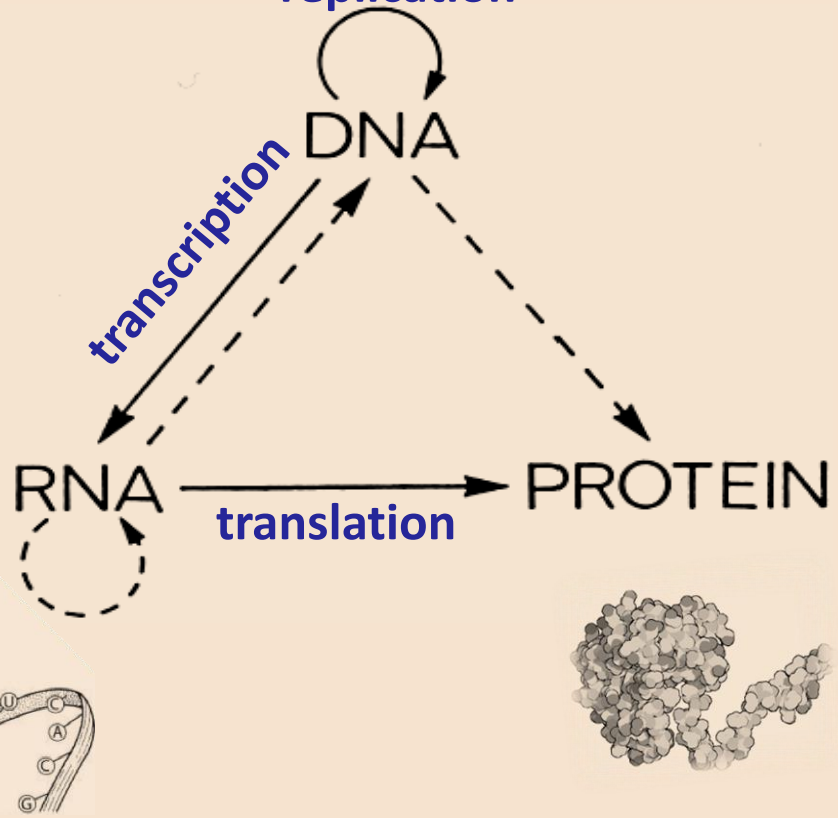
The central dogma graphs the main information channels between nucleotides and proteins

- ***Information from DNA sequence cannot be channeled back from protein to either protein or nucleic acid.***

- 3 information carriers: DNA, RNA protein and 3×3 potential channels:

- 3 general channels (occur in most cells).
- 3 special channels (under “specific” conditions).
- 3 unknown transfers (no example (yet?)).





“special” information transfers

- **Reverse transcription (**RNA → DNA):

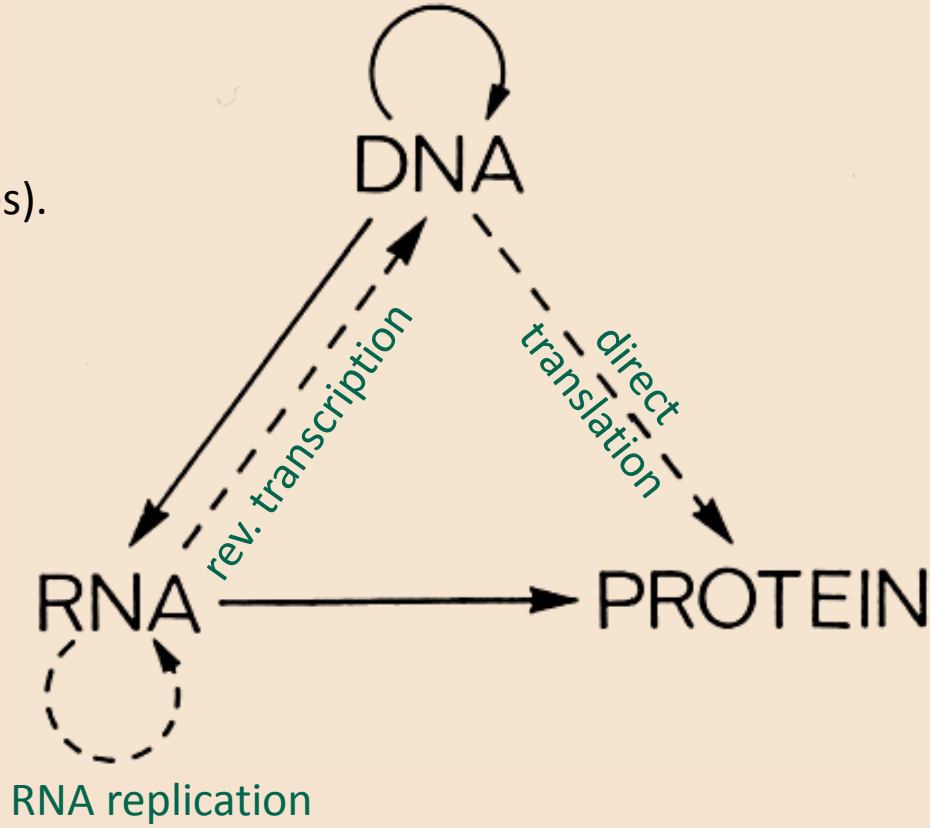
Reverse transcriptase, in retroviruses (e.g. HIV) and eukaryotes (retrotransposons and telomeres).

- **RNA replication** (RNA → RNA):

Many viruses replicate by *RNA-dependent RNA polymerases* (also used in eukaryotes for RNA silencing).

- **Direct translation** (DNA → protein):

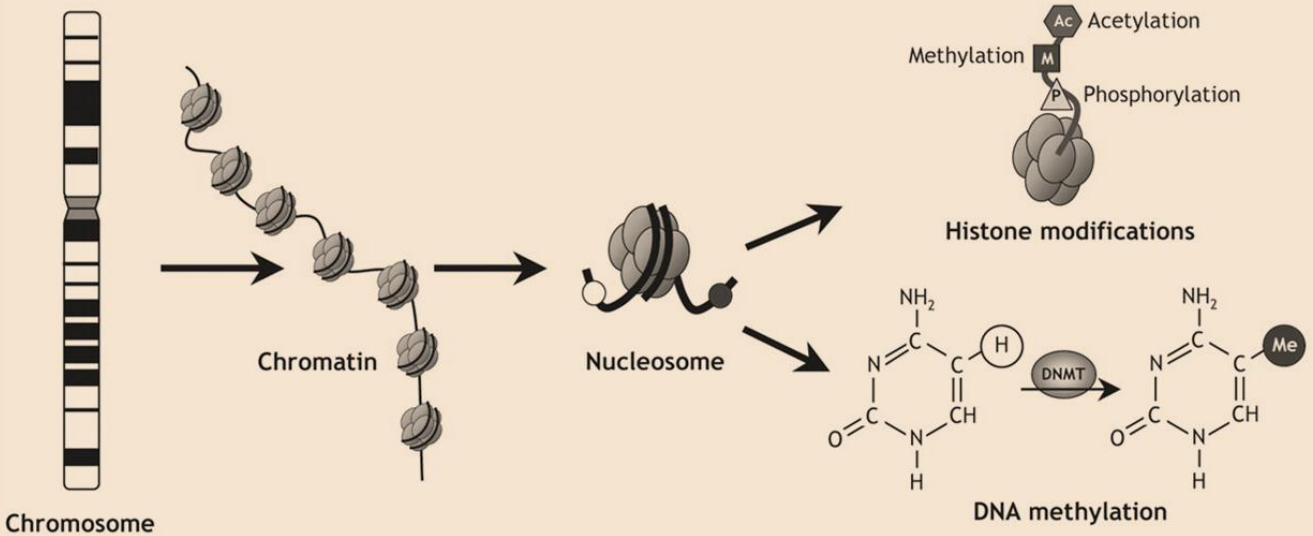
demonstrated in extracts from E. coli which expressed proteins from foreign ssDNA templates.

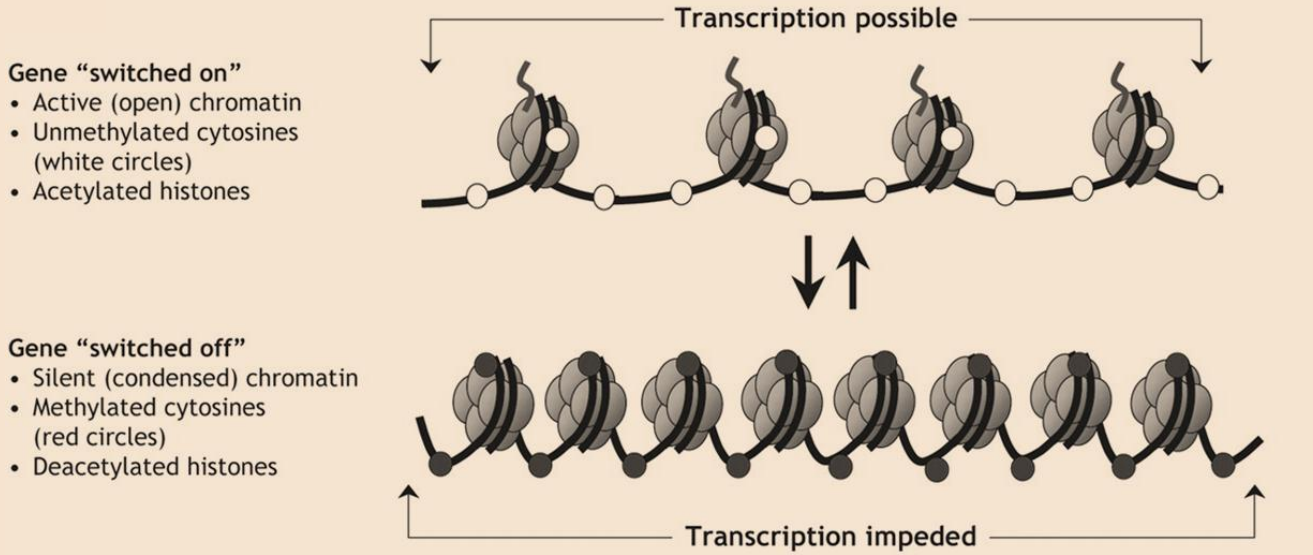


Channels outside the dogma: Epigenetic information transfer

- Changes in methylation of DNA alter gene expression levels.
- Heritable change is called **epigenetic.**
- Effective information change but not DNA sequence.
- Others:

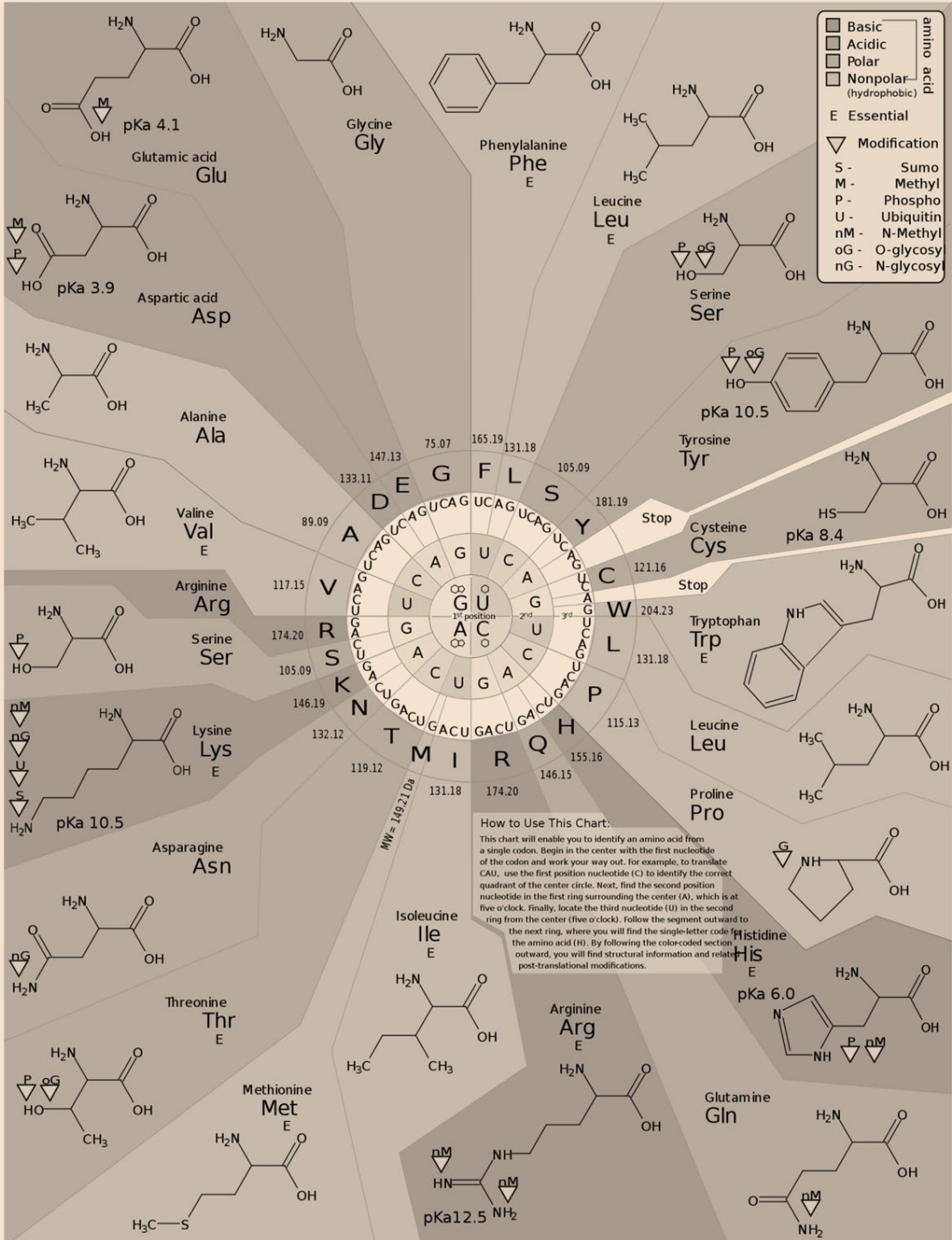
post-translational modification…

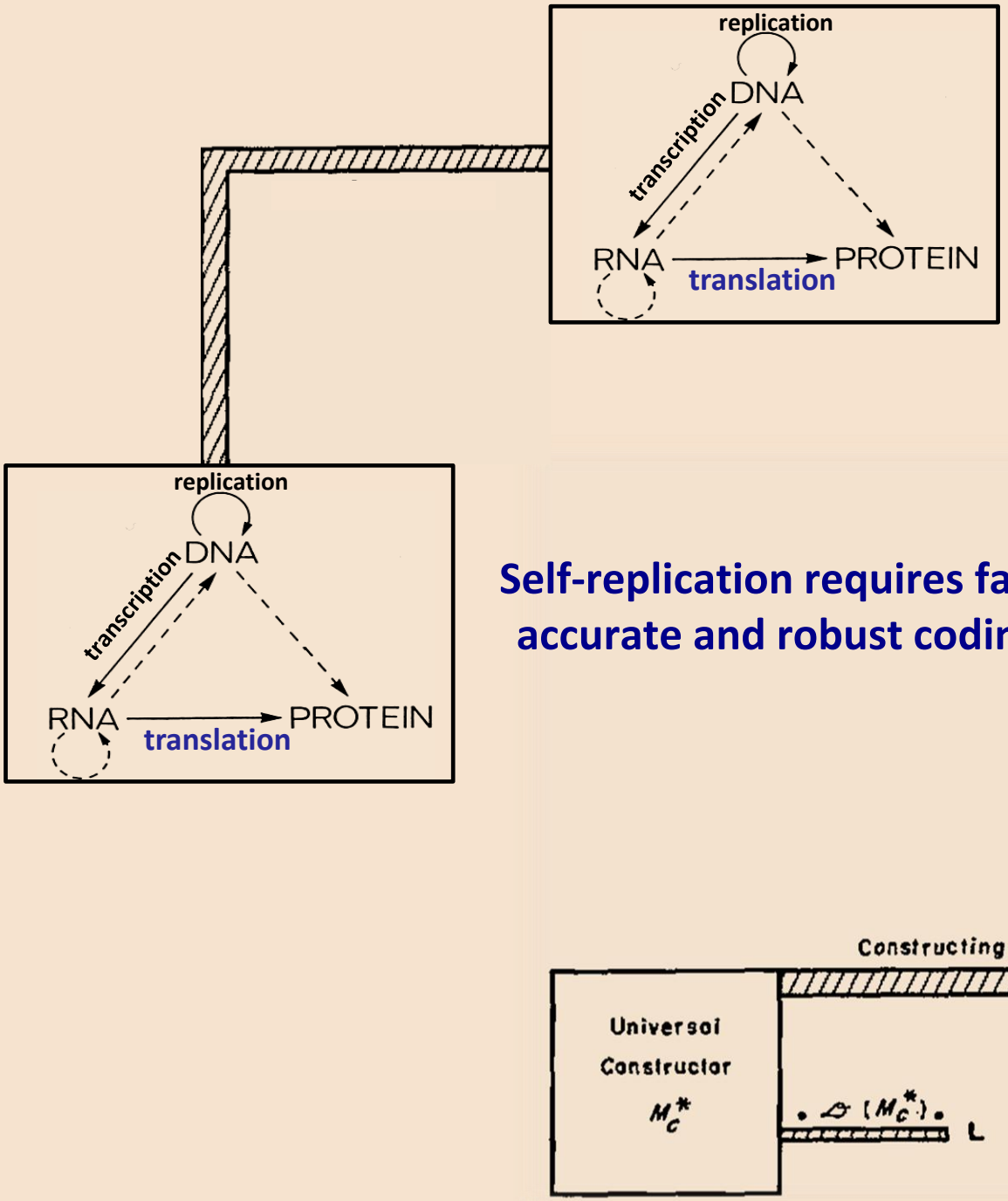




Post-translational modifications of proteins :

- Extends functionality by attaching other groups (e.g. acetate).
- Changes chemical nature of aa.
- Structural changes (disulfide bridges).
- Compensate for missing tRAS (Helicobacter pylori).
- Enzymes may remove amino acids or cut the peptide chain in the middle.

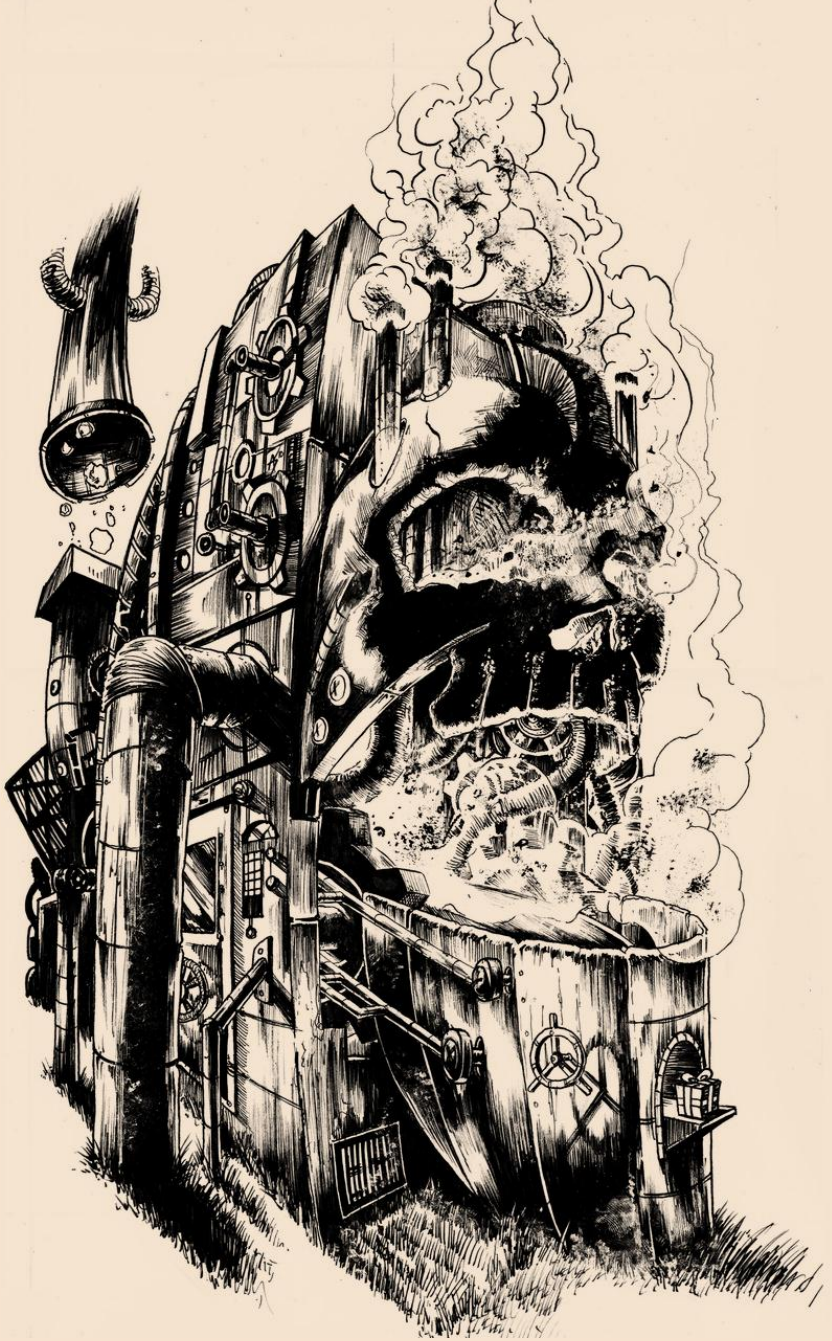




Self-replication requires fast, accurate and robust coding

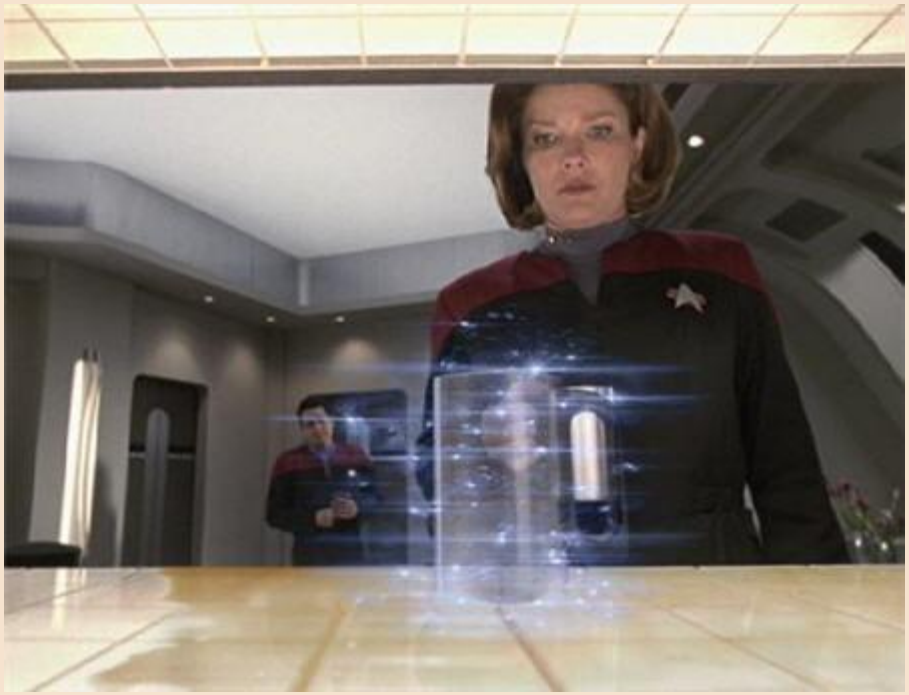
Universal constructors in the arts

The “replicator” (Star Trek)



The Santa Claus machine (A. Sward)

Universal constructors in the arts and in reality (?)

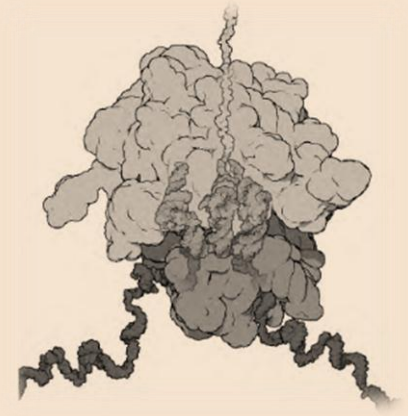


Self-printing?

Outline: molecular codes and errors

- Living and artificial self-replication.
- The main molecular codes of life (central dogma).
- **The translation machinery:**
  - The genetic code, ϕ: codons → amino-acids.
  - The ribosome and the problem of molecular recognition.

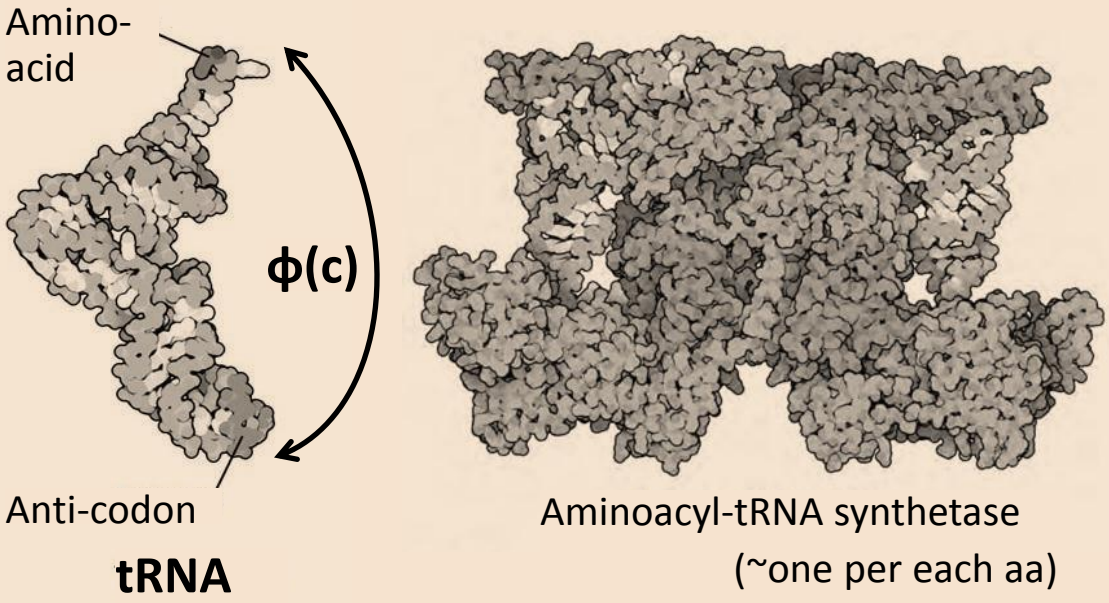
- Basic coding theory: geometrical aspects.
  - How codes cope with errors.
- Emergence and evolution of codes.
- Accuracy vs. rate: proofreading schemes.



The translation machinery is the main system of the living von Neumann's universal constructor

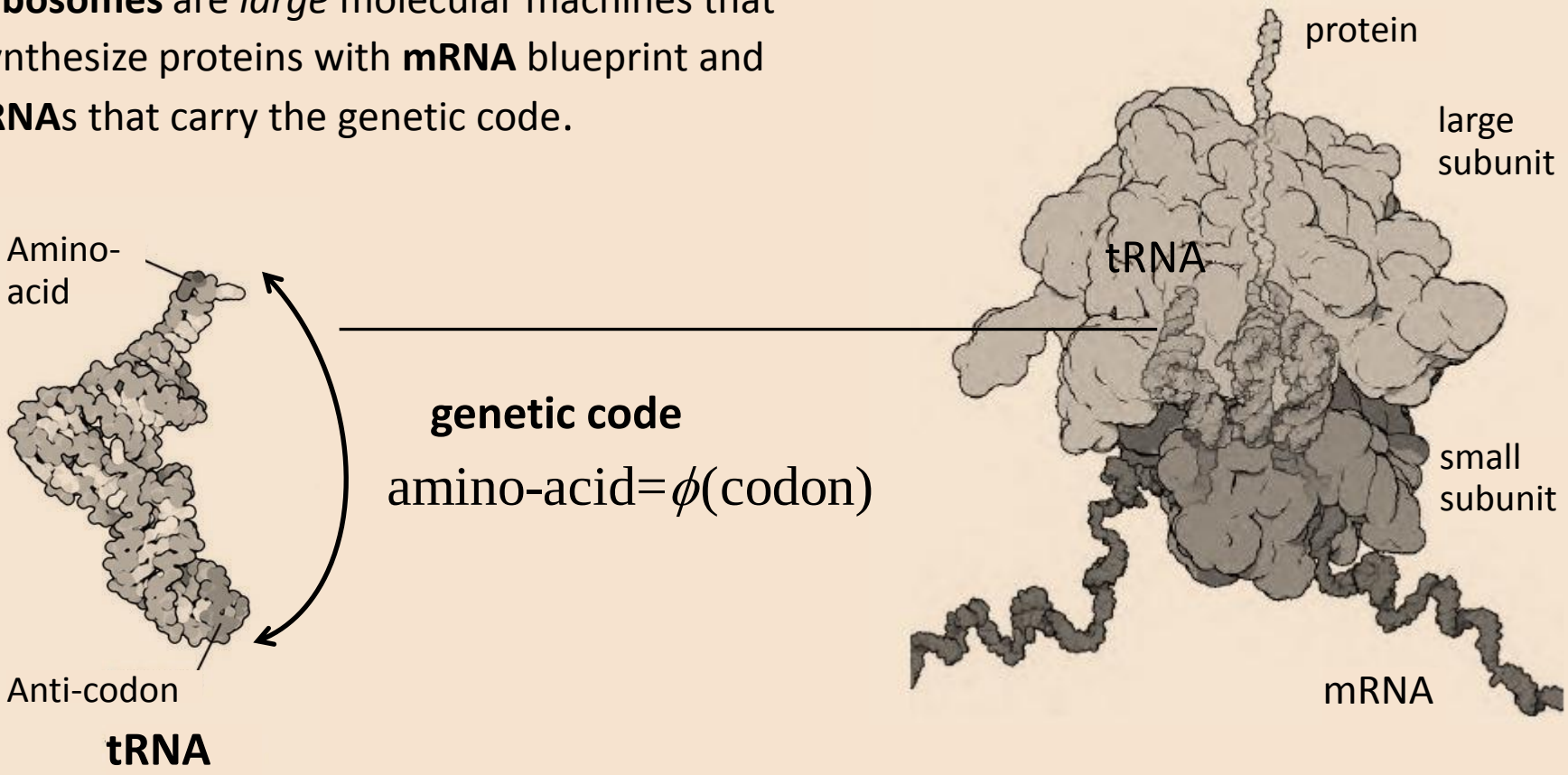
- Machinery parts = **tRNA** + **synthetase** + **ribosome…**
- The translation machinery conveys information from nucleotides to proteins.

- Synthetases charge tRNAs according to the genetic code.



Ribosomes translate nucleic bases to amino acids

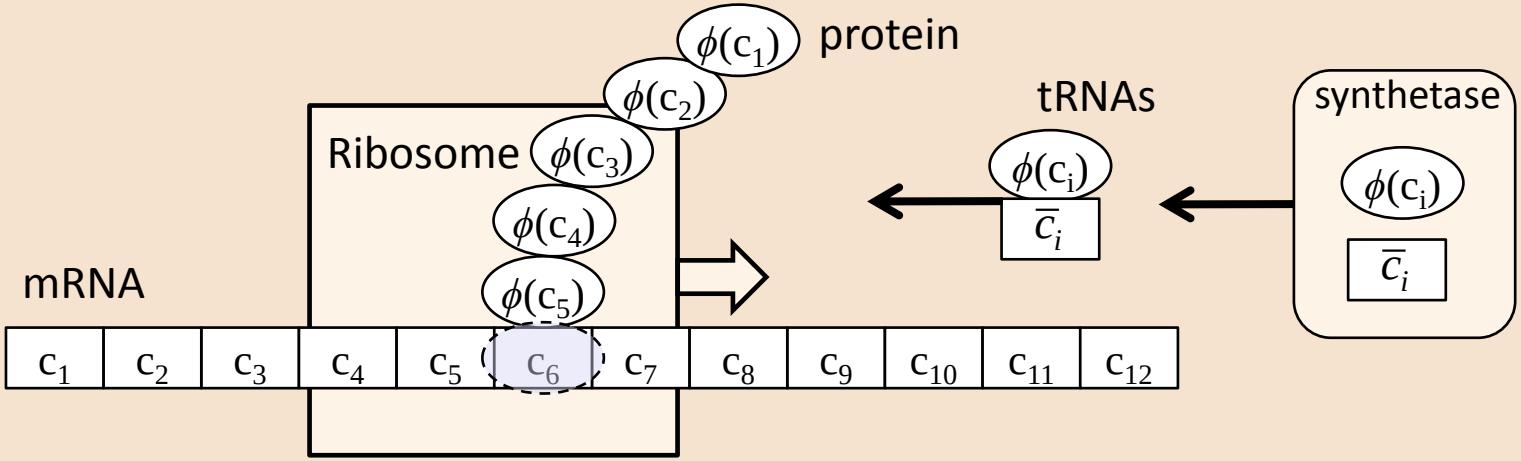
- **Ribosomes** are *large* molecular machines that synthesize proteins with **mRNA** blueprint and **tRNA**s that carry the genetic code.

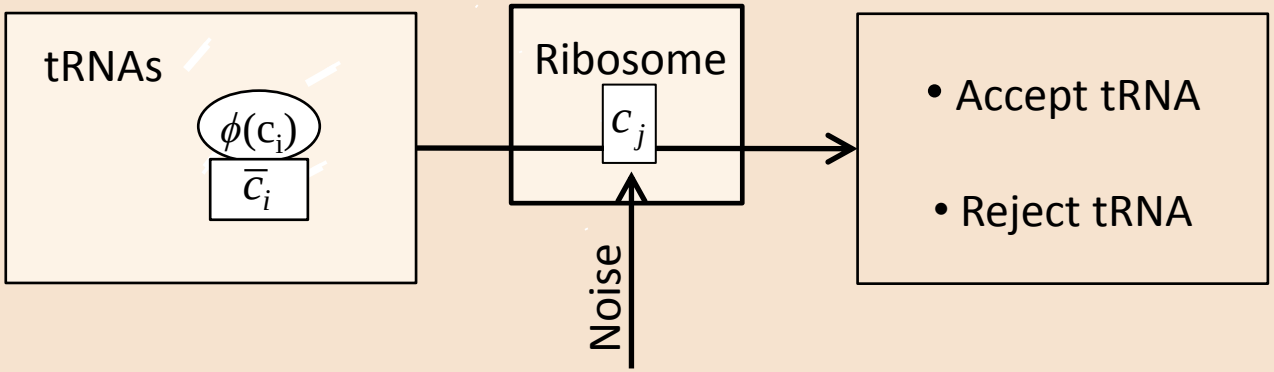


Goodsell, *The Machinery of Life*

1. Is the code ϕ(c) adapted to the noise problem?

Ribosome needs to recognize the correct tRNA



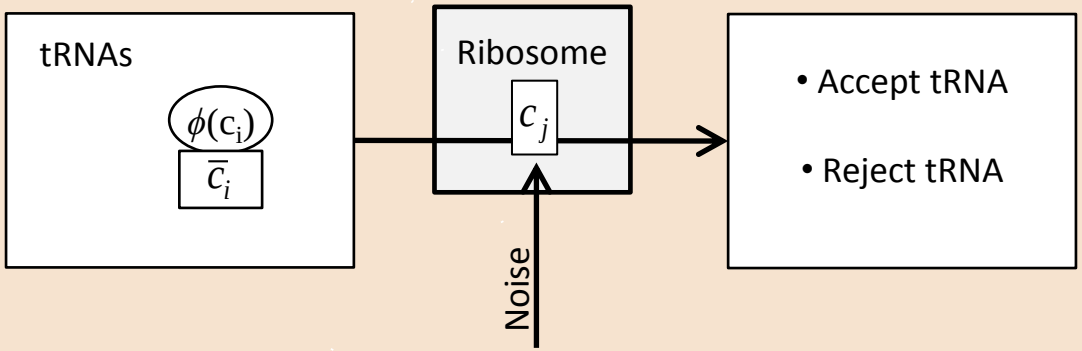


(i) binding wrong tRNAs: $\text{amino-acid} \neq \phi(\text{codon})$

(ii) unbinding correct tRNAs: $\text{amino-acid} = \phi(\text{codon})$

2. How to construct fast\accurate\small ***molecular*** decoder ?

2.Decoding at the ribosome is a molecular recognition problem



- Central problem in biology and chemistry:

How to evolve molecules that recognize in a noisy environment?

(crowded, thermally fluctuating, weak interactions).

- How to estimate recognition performance ("fitness")?
- What are the relevant degrees-of-freedom? **Dimension**? **Scaling**?
- What is the role of conformational changes?

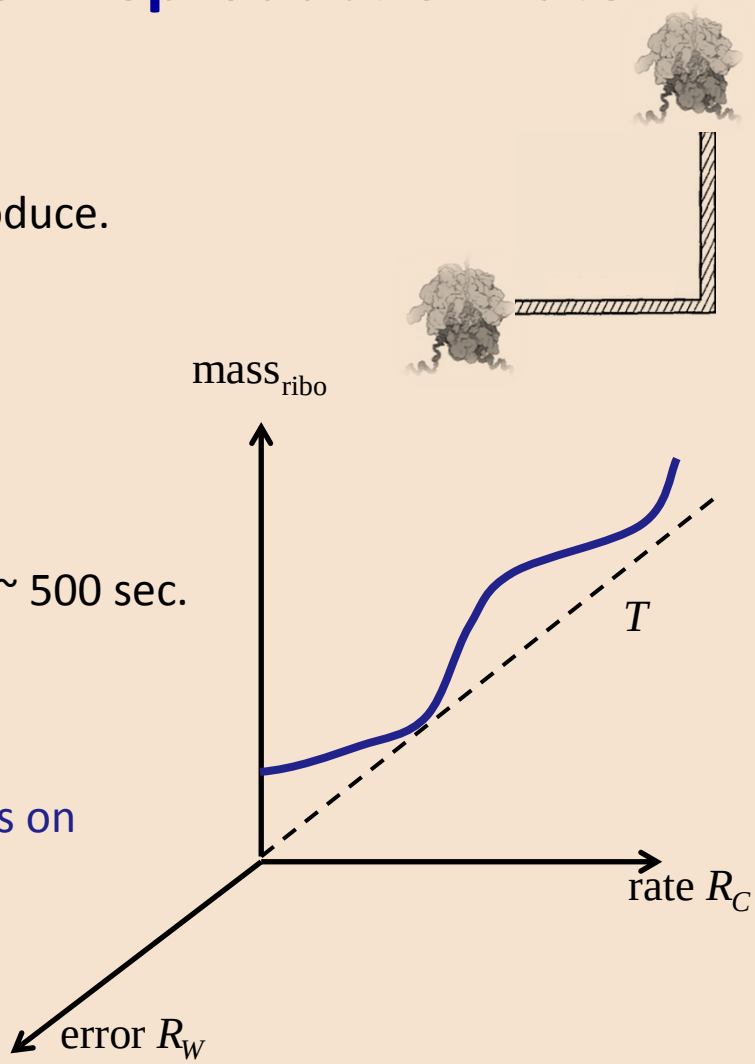
Ribosome sets physical limit on self-reproduction rate

Large fraction of cell mass is ribosomes.

- In self-reproduction each ribosome should self-reproduce.
- Sets lower bound on self-reproduction rate .

$$T \geq \frac{\text{mass}_{\text{ribo}}}{R_C} \approx \frac{10^4 \text{ amino-acids}}{20 \text{ amino-acids/sec}} = 500 \text{ sec}$$

- "Fastest " growing bacteria (*Clostridium perfringens*): T ~ 500 sec.



Problem: how ribosome accuracy affects fitness depends on

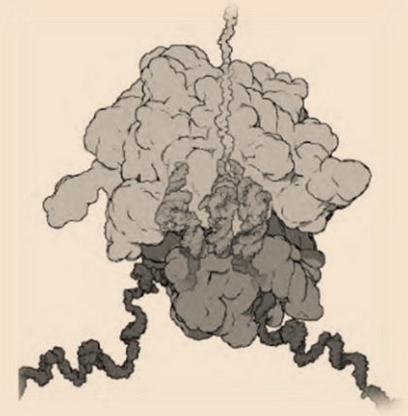
(i) Basic protein properties (mutations).

(ii) Biological context (environment etc.).

Outline: molecular codes and errors

- Living and artificial self-replication.
- The main molecular codes of life (central dogma).
- **The translation machinery:**
  - The genetic code, ϕ: codons → amino-acids.
  - The ribosome and the problem of molecular recognition.

- Basic coding theory: geometrical aspects.
  - How codes cope with errors.
- Emergence and evolution of codes.
- Accuracy vs. rate: proofreading schemes.

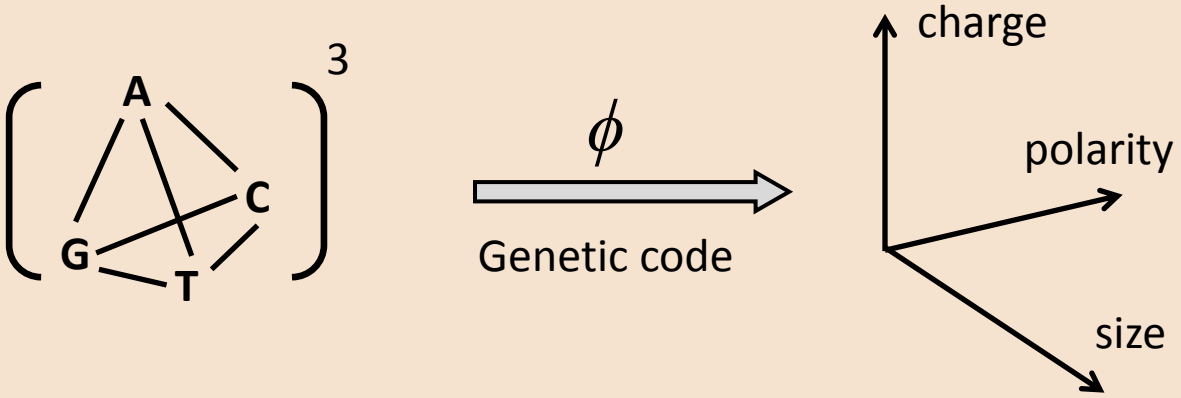


1. The genetic code maps DNA to protein

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

$$\text{codon} = b_1 b_2 b_3,\ b_i \in \{\text{A, T, G, C}\}.$$
$$\phi(\text{codon}) \rightarrow \text{amino-acid}.$$

- Genetic code embeds the codon-graph (Hamming graph) into space of amino-acids ("digital to analog").



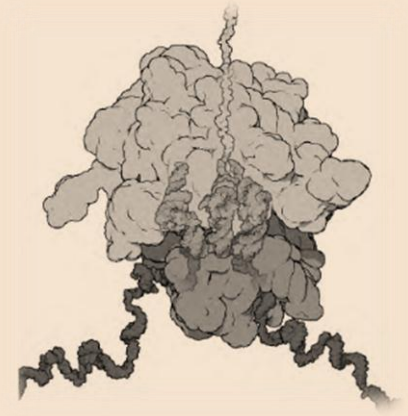
- **Translation machinery**, whose main component is the **ribosome**, facilitates the map.

Outline: molecular codes and errors

- Living and artificial self-replication.
- The main molecular codes of life (central dogma).
- The translation machinery:
  - The genetic code, ϕ: codons → amino-acids.
  - The ribosome and the problem of molecular recognition.

⇨

- **Basic coding theory: geometrical aspects.**
  - How codes cope with errors.
- Emergence and evolution of codes.
- Accuracy vs. rate: proofreading schemes.



Coping with unreliability of coding machinery

- *Error detecting code* - parity checking.
- One check: Odd parity → mistake (e.g. 0111).
- Retransmission.

- Single error can be detected but not corrected.

- The redundancy of the code:

$$R = \frac{\text{total \# of bits}}{\text{\# of message bits}} = \frac{n}{n-1}$$

signal	binary	parity
0	000	0
1	001	1
2	010	1
3	011	0
4	100	1
5	101	0
6	110	0
7	111	1

Error correction requires minimal redundancy

- *Error correcting code* – can detect and correct errors.
- Multiple checks – Locating errors by confluence.

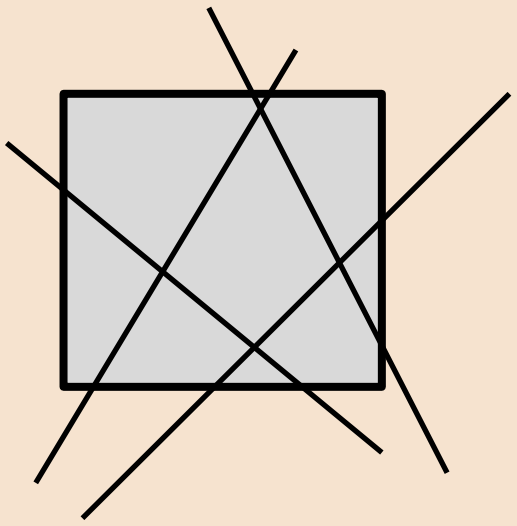
	1001010111
0	1001101101
1	0010011110
0	0011111001
0	1101010001
1	1101001100

- *Triplication* code – send each message thrice (R = 3).

Rectangular code $R = 1 + \frac{2}{\sqrt{n}}$

- What is the minimal number of checks *m*?
  - To locate *n* positions requires $2^m \geq n+1.$

$$R = \frac{n}{n-m} \geq \left(1 - \frac{\log_2(n+1)}{n}\right)^{-1} \simeq 1 + \frac{\log_2 n}{n}$$

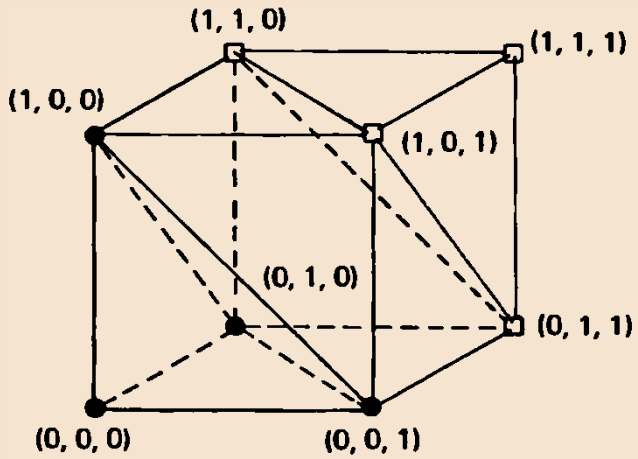


- Hamming's code reaches this limit.

Geometric view of error correction and detection

- Messages are mapped between hypercubes

$$\phi : Y_n \to Y_{n+m} \quad (1001 \to 1001101)$$

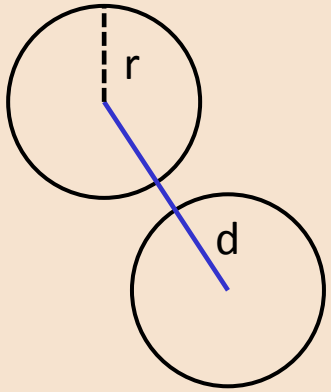


1 Three-dimensional spheres about (0, 0, 0) and (1, 1, 1)

- Metric is the Hamming distance:

$$\left| x_i - x_j \right| = \#\ \text{of different letters}$$

- Sphere is $S = \{x \mid \ | x - x_0 | \le r\}.$

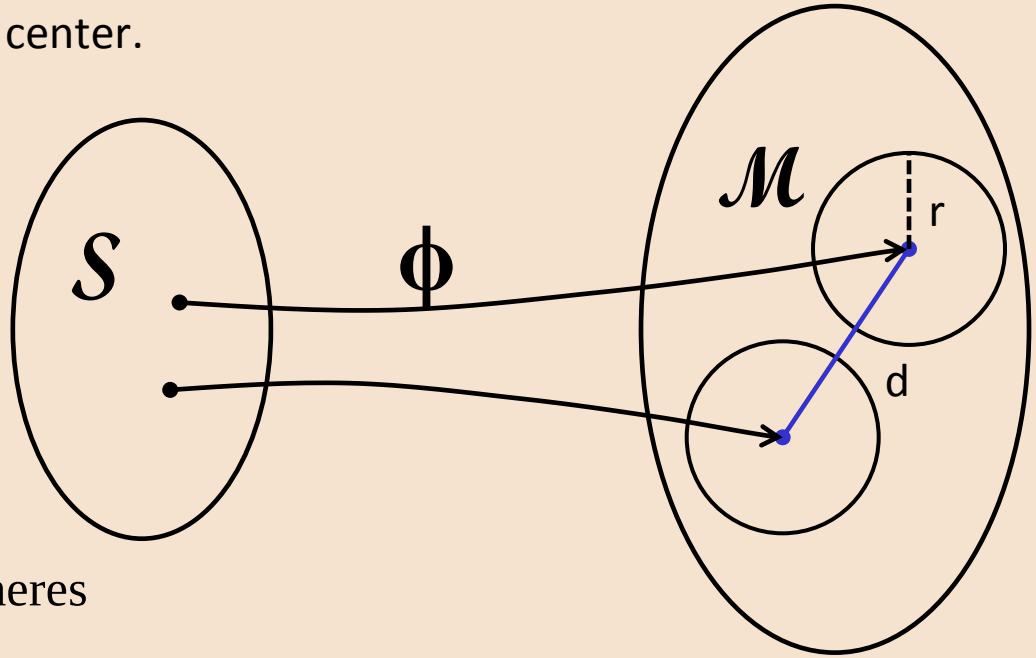


Error correction is packing hard spheres

- To correct *r* errors the spheres should be at least at distance $d = 2r + 1.$
- Correction: move to nearest sphere center.

- How many words can be encoded?

Or how many spheres can be packed?



$$\text{total volume} \geq \text{sphere volume} \times \#\,\text{spheres}$$

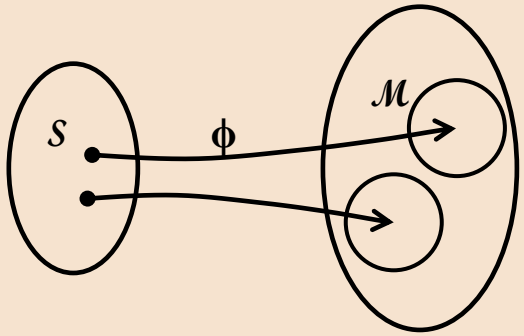
$$2^n \geq (n+1) \times 2^{n-m} \Rightarrow 2^m \geq n+1$$

$$\#\,\text{spheres} = \#\,\text{words} = 2^{n-m} \leq \frac{2^n}{n+1}$$

Shannon's channel coding theorem sets upper limit on the capacity of a noisy channel

- Noisy channel is defined by stochastic input\output $\phi(s|m)$.
- **Channel capacity** measures the input\output correlation

$$C = \max_{\phi(s)} I(S;M) = \max_{\phi(s)} \left\langle \log_2 \frac{\phi(s,m)}{\phi(s)\phi(m)} \right\rangle.$$

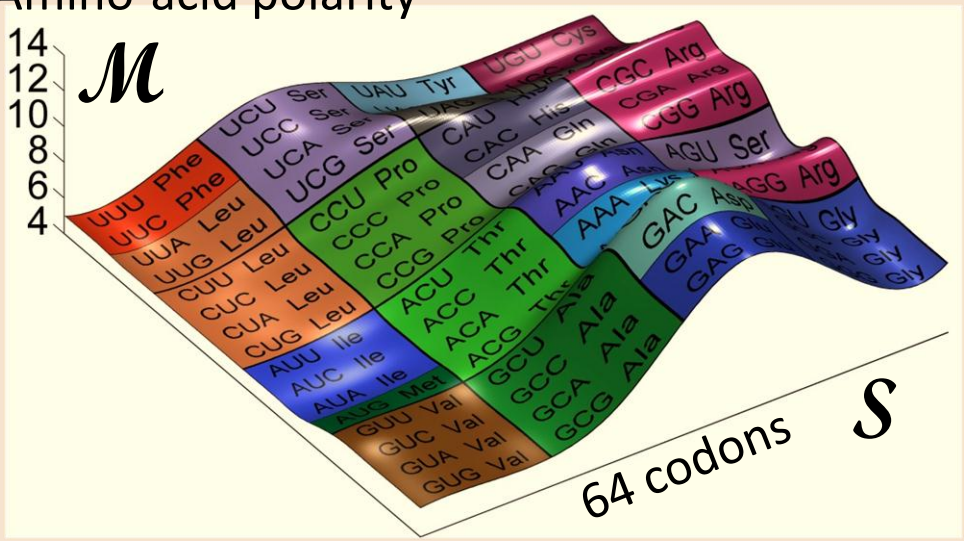


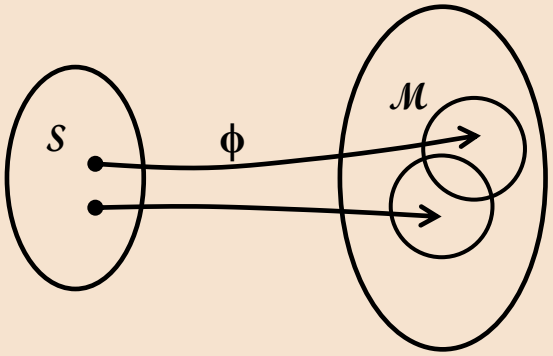
- **Channel rate** $R = \lim_{n\to\infty} \frac{\log_2(\#\text{ words})}{n}.$
- Shannon's coding theorem (1948/9): $R = C.$
- Proof: show that # hard spheres is $2^{nR} = 2^{nI(S,M)} = 2^{nC}.$
- Upper limit achieved only "recently" (turbo codes, LDPC).

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

Amino-acid polarity





- Degenerate (20 out of 64) → “spheres”.
- Compactness of amino-acid regions.
- **Smooth** (similar “color” of neighbors).
- But not immune to one-letter errors (“soft” spheres).

Generic properties of molecular codes?

Gray code ”smooths” the impact of errors

- Invented by Émile Baudot for telegraphy.
- Often used in A→D and D→A applications.
- Minimizes the number of changes between close by values → smooth code.
- Used in many modulation schemes.

(e.g. phase shift).

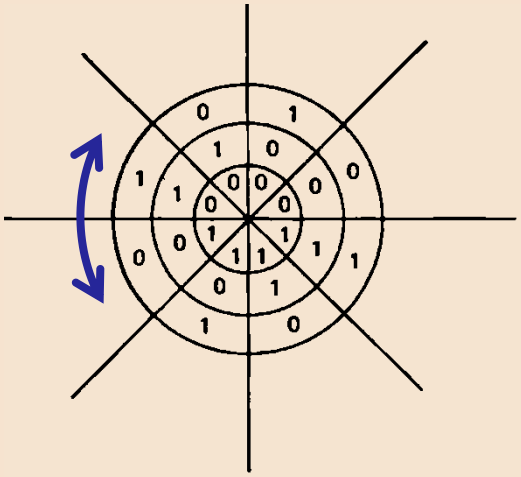
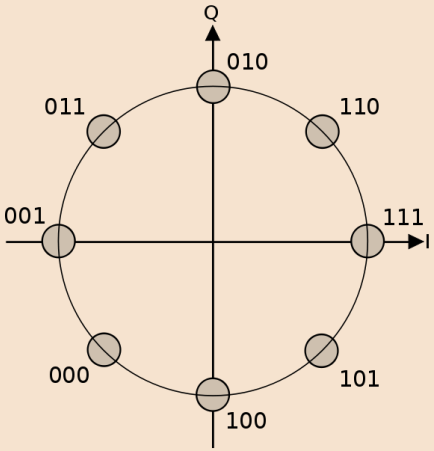
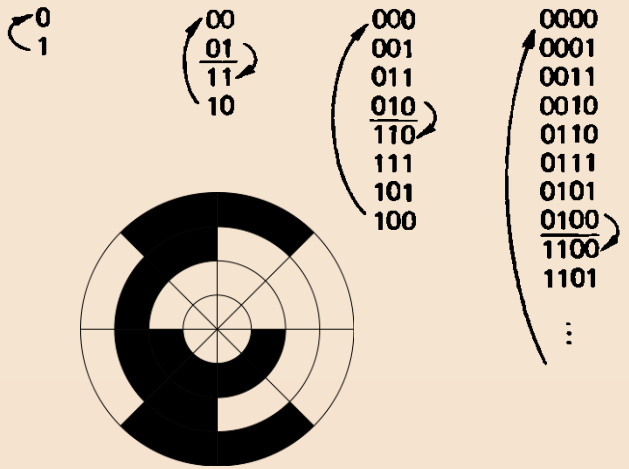


Figure 5-15.1 Eight-sector binary wheel





The smooth genetic code as a combinatorial game

<u>"Marble packing"</u>

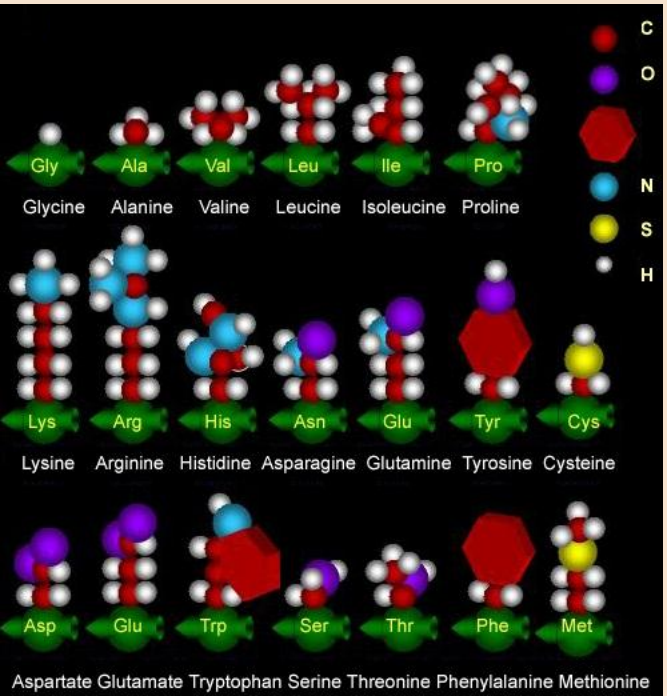
(A) Max colors.

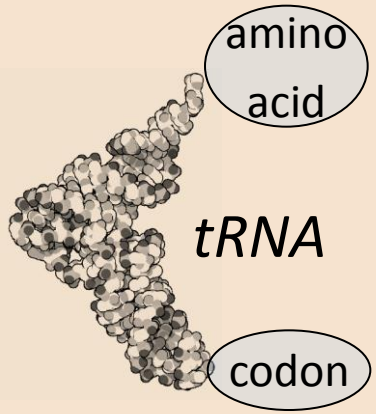
(B) Same\similar color of neighbors.

The genetic code maps codons to amino-acids

- Molecular code = map relating two sets of molecules.
- Spaces defined by similarity of molecules (size, polarity etc.)

20 amino-acids



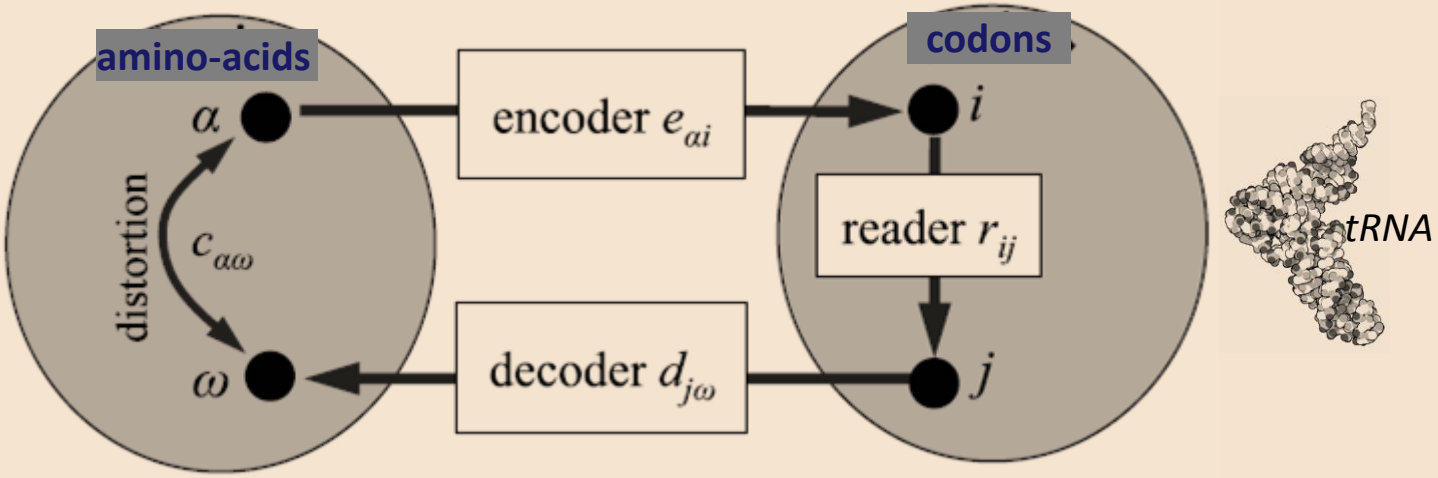


Genetic Code

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

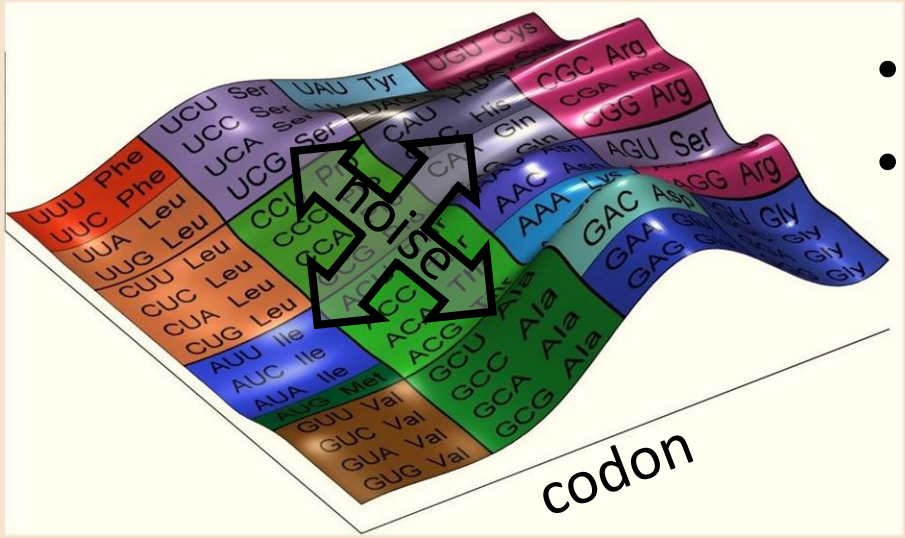
Fitter codes have minimal distortion



$$\mathrm{D} = \langle c_{\alpha\omega} \rangle = \sum_{paths} p_{path} c_{\alpha\omega} = \mathrm{Tr}\left(e \cdot r \cdot d \cdot c\right)$$

- Distortion of noisy channel, ***D*** = average distortion of AA.
- ***r*** defines topology of codon space.
- ***c*** defines topology of amino-acid space.

Smooth codes minimize distortion

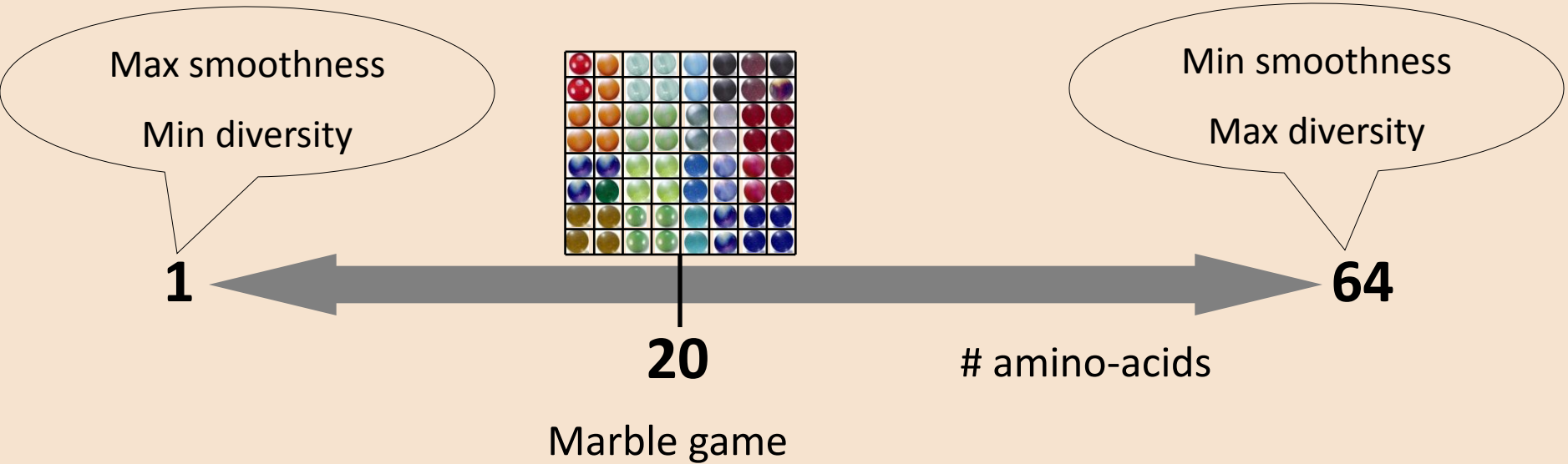


- Noise confuses close codons.
- Smooth code:

  close codons = close amino-acids.

  → minimal distortion.

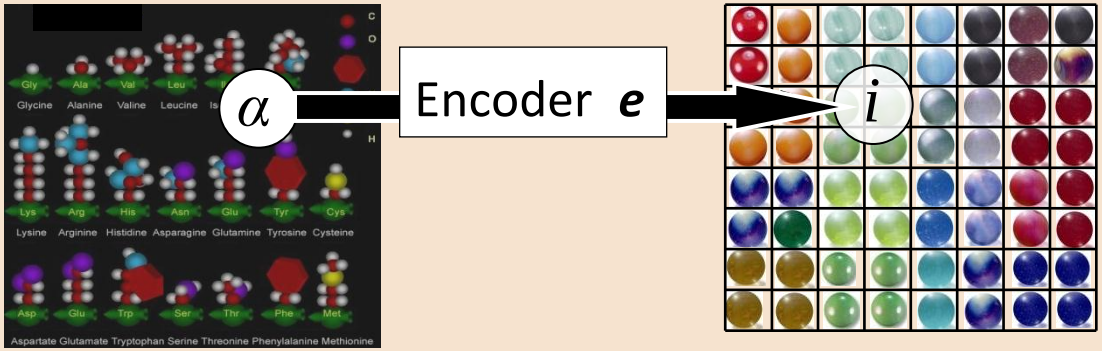
- Optimal code must balance contradicting needs for **smoothness** and **diversity**.



Channel rate is code's cost

- Diverse codes require high **specificity** = high binding energies ε.
- Cost ~ average binding energy $< \varepsilon >$.
- Binding prob. ~ Boltzmann: $\boldsymbol{E} \sim e^{\varepsilon/T}$.

$$I \sim \sum_{\alpha,i} e_{\alpha i} \ln e_{\alpha i} \sim \left\langle \varepsilon_{\alpha i} \right\rangle_e$$

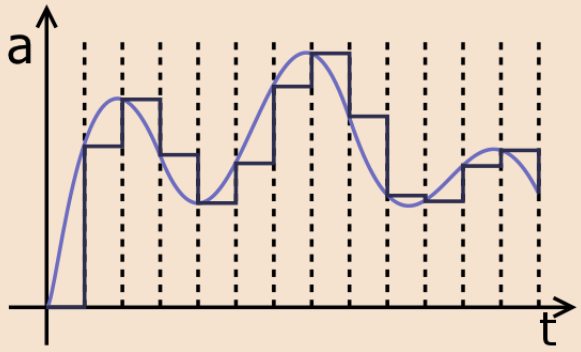


- Cost ***I*** = **Channel Rate** (bits/message)

Rate-distortion theory of noisy information channels

- How well a mapping represents a signal?
- Example: quantization of continuous signal.
- The average distortion of a signal

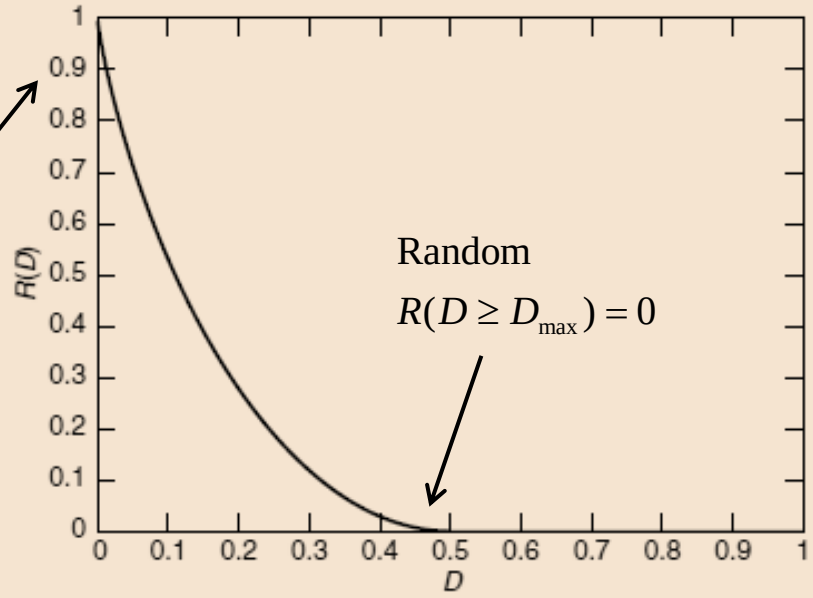
$$\mathrm{D} = \langle c_{\alpha\omega} \rangle = \sum_{paths} p_{path} c_{\alpha\omega}$$



- Main theorem: there exists a *rate-distortion function* $R(D)$ which is the minimal required rate R to achieve distortion D.

$$\mathrm{R(D)} = \min_{D\{\phi\} \le D} I(S, M)$$

Shannon's limit:
$R(D = 0) = C$



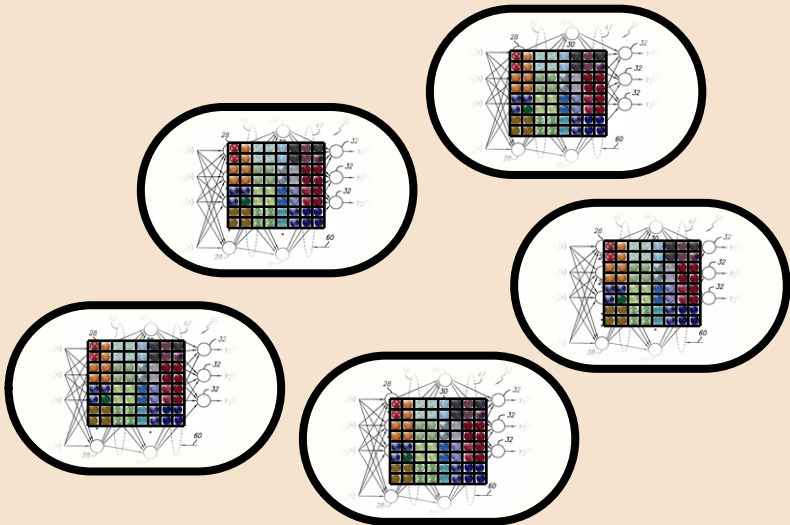
(Shannon, Kolmogorov 1956)

Code's fitness combines rate and distortion of map

Fitness = Gain x *Distortion + Rate* $\qquad H = \kappa D + I$

- ***Gain*** β increases with organism complexity and environment richness.
- Fitness ***H*** is "free energy" with inverse "temperature" κ.
- Evolution varies the gain κ.

- Population of self-replicators evolving according to code fitness *H*: mutation, selection, random drift.



Code emerges at a critical coding transition

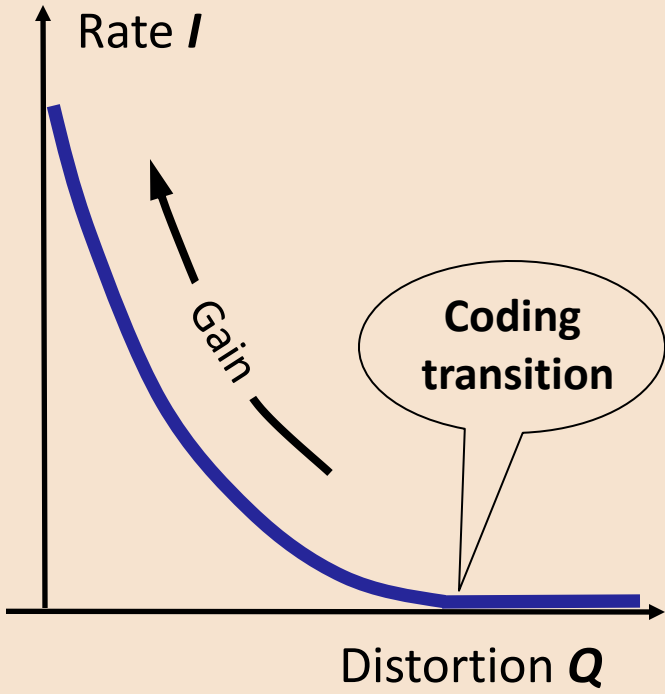
- Low gain β : Cost too high

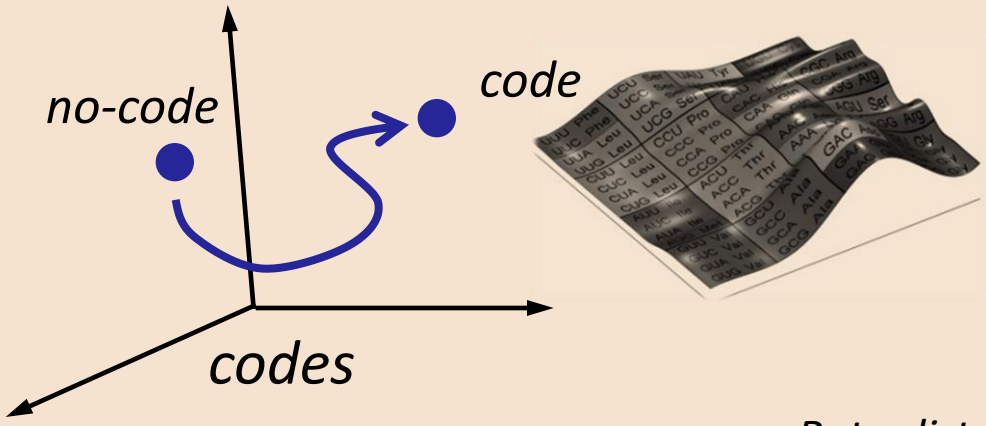
  → no specificity → **no code**.

- **Code emerges** when β increases:

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

- Continuous phase transition.
- Emergent code is smooth, low mode of ***R***.

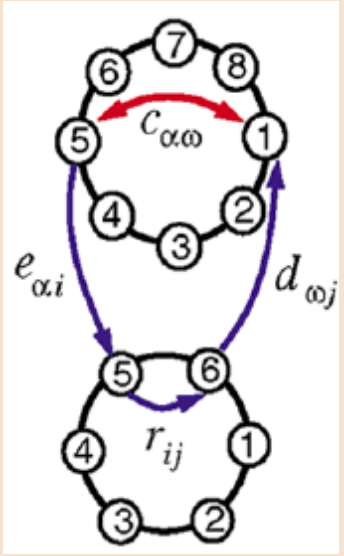


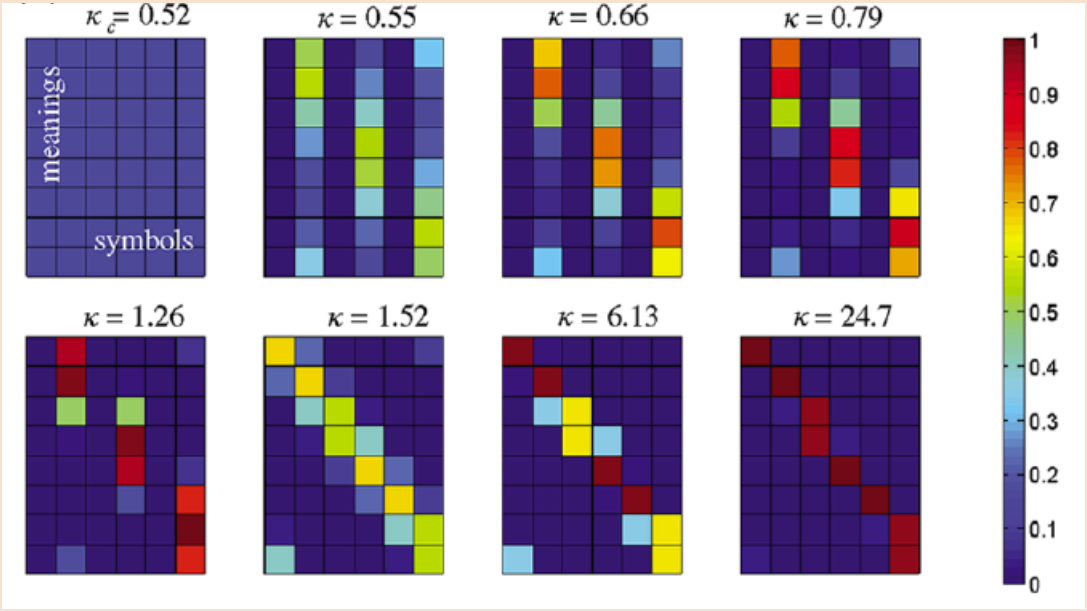


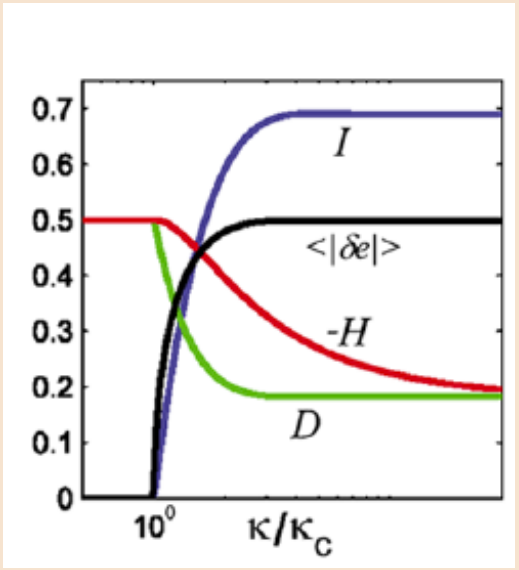
Rate-distortion theory (Shannon 1956)

The emergent code is smooth

- Example: mapping between two cycles.
- Code emerges at critical transition.





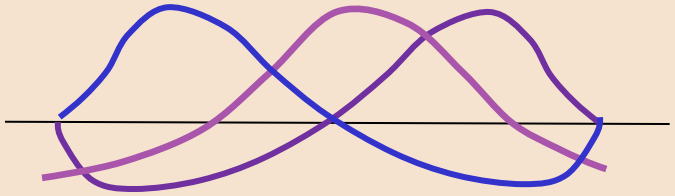


- Order parameter: deviation from random map

$$\delta e_{\alpha i} = e_{\alpha i} - e_{ai}^{\text{rand}}.$$

PRL 2008

Emergent code is a smooth mode of error-Laplacian

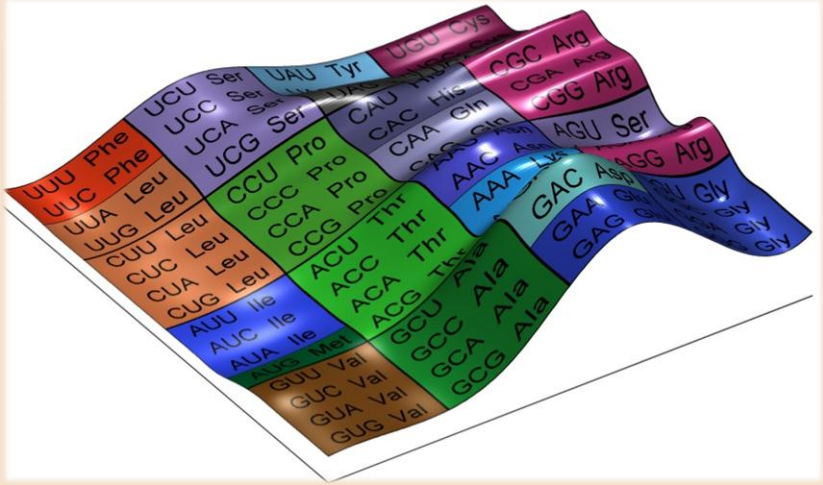


- Lowest excited modes of graph-Laplacian *R* .
- Single maximum for lowest excited modes (Courant).
- Every mode corresponds to amino-acid :

 # low modes = # amino-acids.

→ single contiguous domain for each amino-acid.

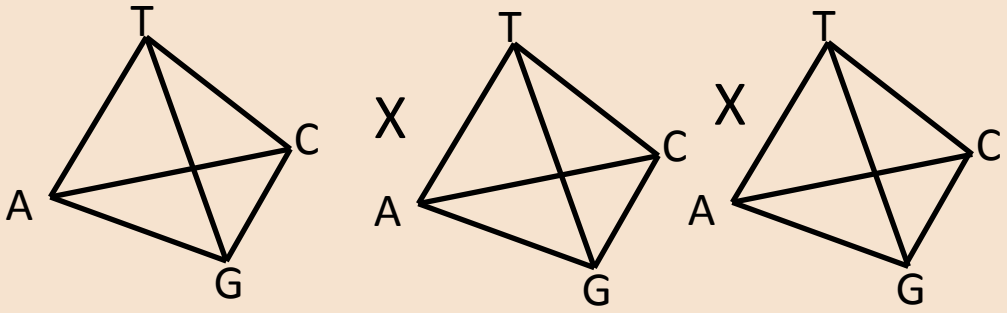
→ **Smoothness**.

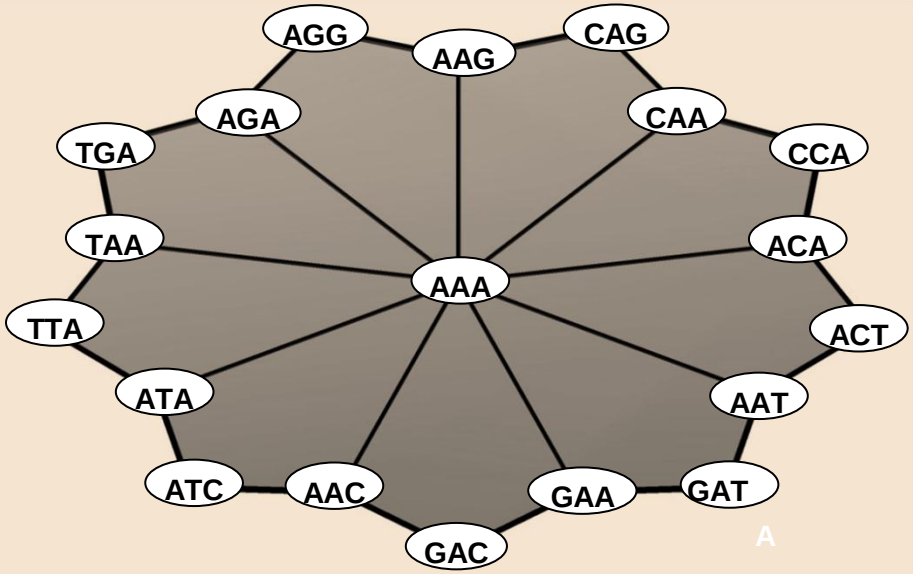


Probable errors define the graph and the topology of the genetic code

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

$K_4 \times K_4 \times K_4$





- Non-planar graph (many crossings).
- Genus γ = # holes of embedding manifold.
- Graph is holey : embedded in $\gamma = 41$

(lower limit is $\gamma = 25$)

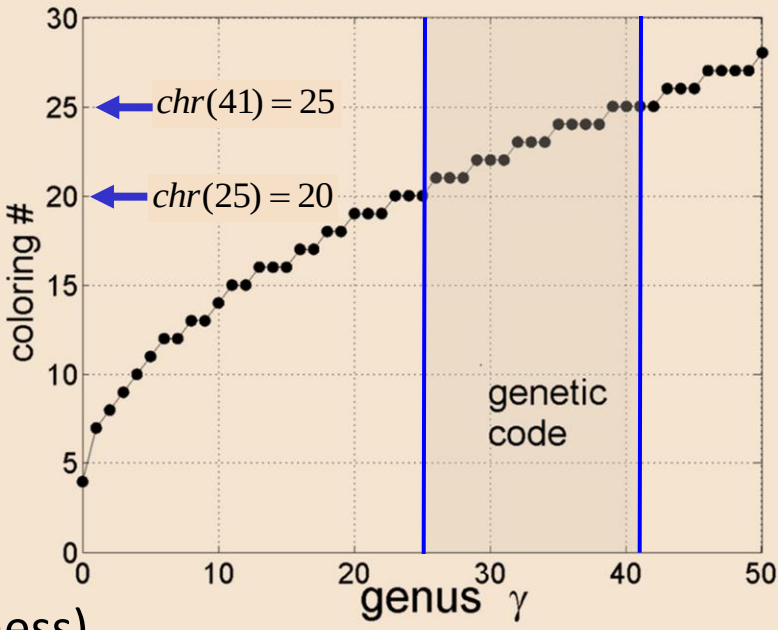
Coloring number limits number of amino-acids

- Q: Minimal # colors suffices to color a map where neighboring countries have different colors?
- A: Coloring number, a topological invariant (function of genus):

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

(Ringel & Youngs 1968)

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

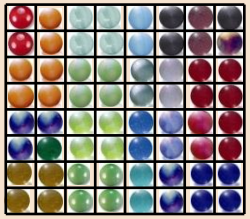


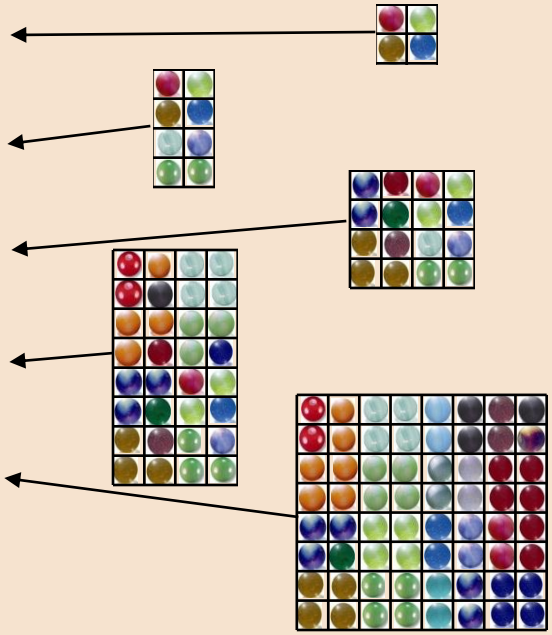
- From Courant 's theorem + "convexity" (tightness).
- Genetic code: γ = 25-41 → coloring number = 20-25 amino-acids

The genetic code coevolves with accuracy

- A path for evolution of codes: from early codes with higher codon degeneracy and fewer amino acids to lower degeneracy codes with more amino acids.

1st	2nd	3rd	γ	*chr #*
1	4	1	0	4
2	4	1	1	7
4	4	1	5	11
4	4	2	13	16
4	4	3	25	20
4	4	4	41	25





Part I: Summary

- The translation machinery:
  - The genetic code, ϕ: codons → amino-acids.
- Genetic code is a smooth map that minimizes distortion.
- Model for emergence: phase transition in a noisy mapping.
- Free energy is rate-distortion function.
- Continuous transition.
- Topology governs emergent code.

Sources:

- Shannon, Mathematical Theory of Communication.
- Hamming, Coding and computation.
- von Neumann, In Automata Studies.
- Feynman, Lectures on Computation.
- Cover & Thomas, Elements of information theory.
- Berger T, Rate distortion theory.

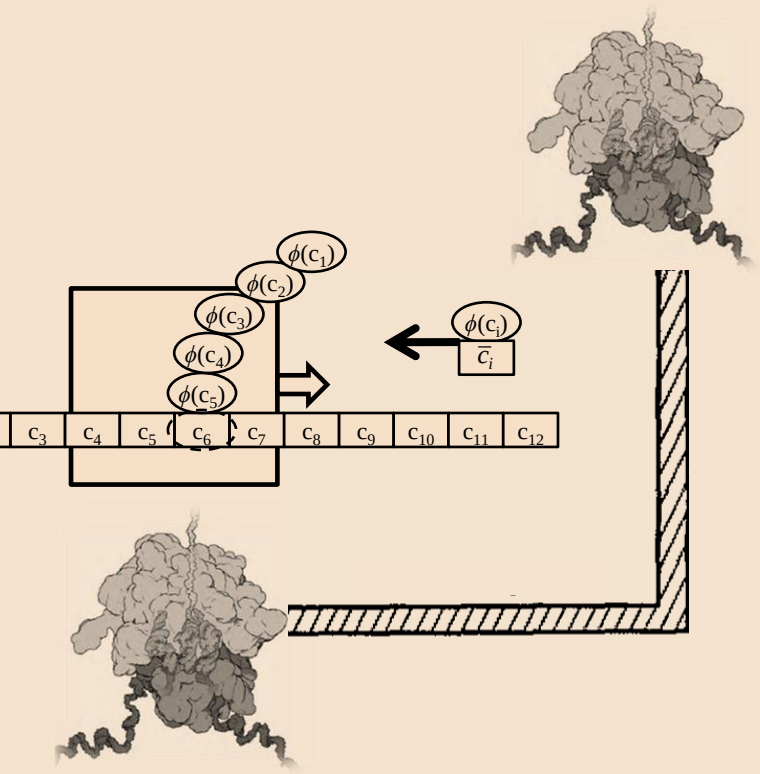
Papers on coding: follow PITP link

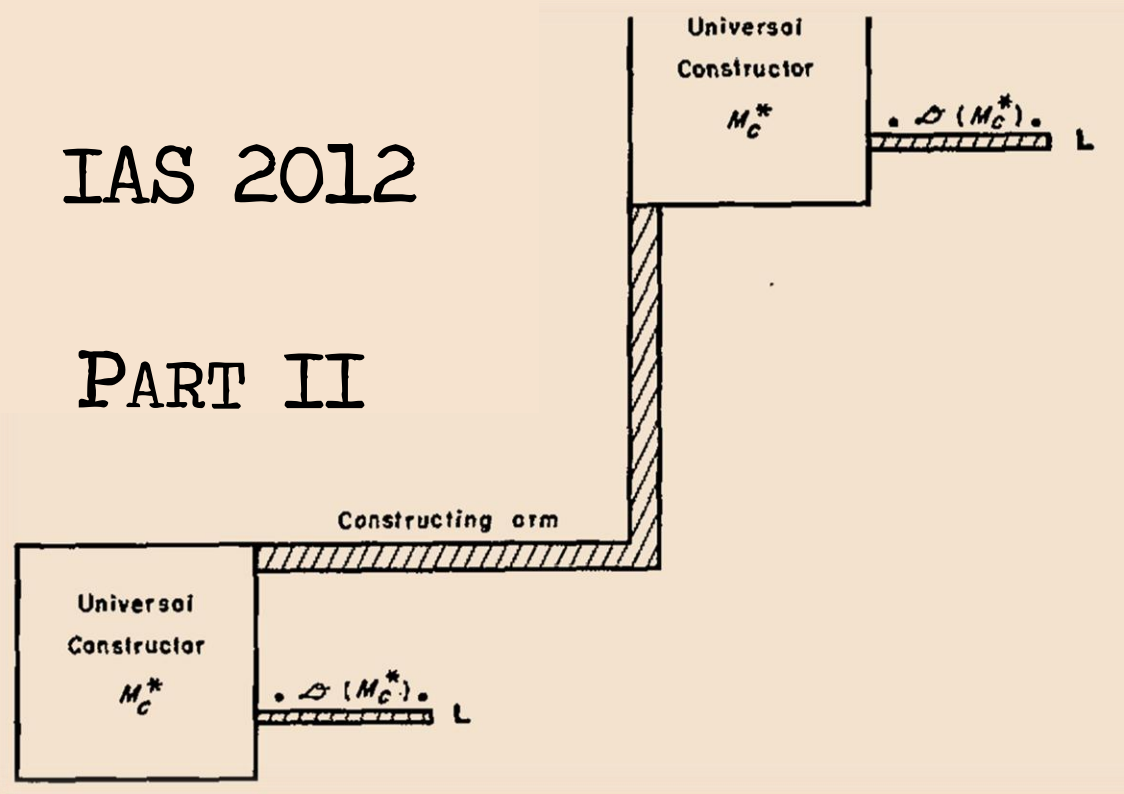
Coding a Life Full of Errors

PITP

IAS 2012

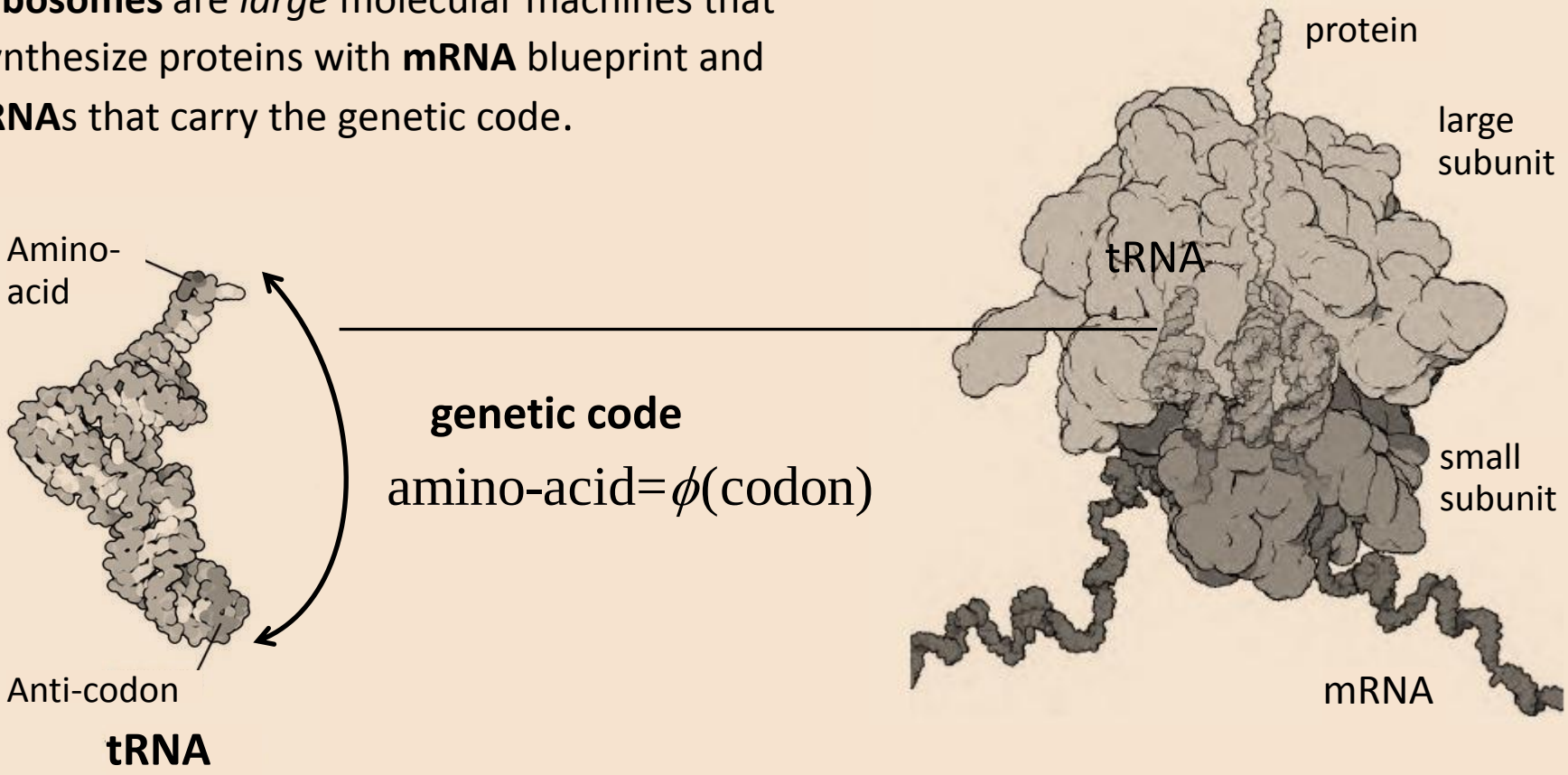
Part II





Ribosomes translate nucleic bases to amino acids

- **Ribosomes** are *large* molecular machines that synthesize proteins with **mRNA** blueprint and **tRNA**s that carry the genetic code.



Goodsell, *The Machinery of Life*

Is the code ϕ(c) adapted to the noise problem?

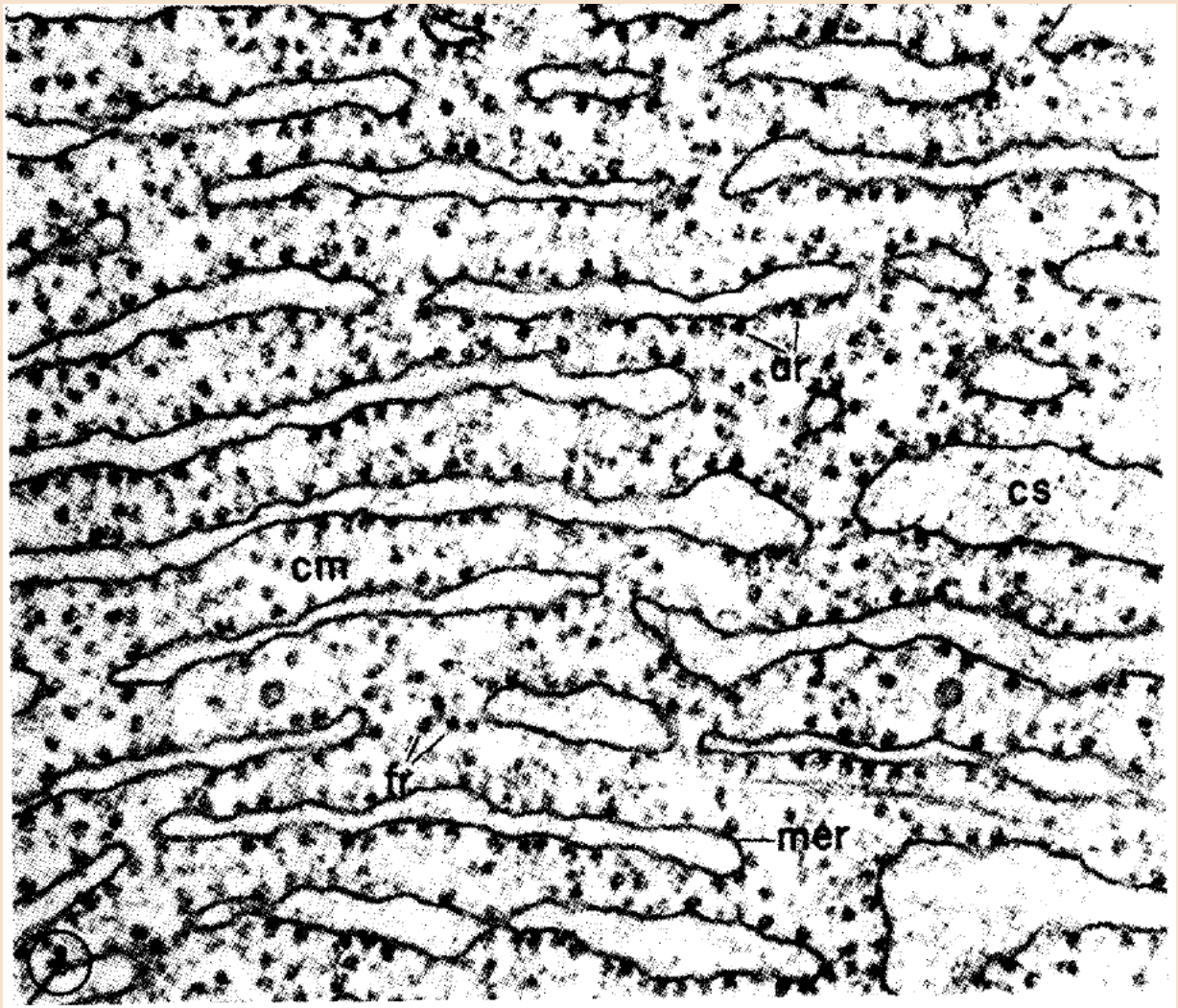


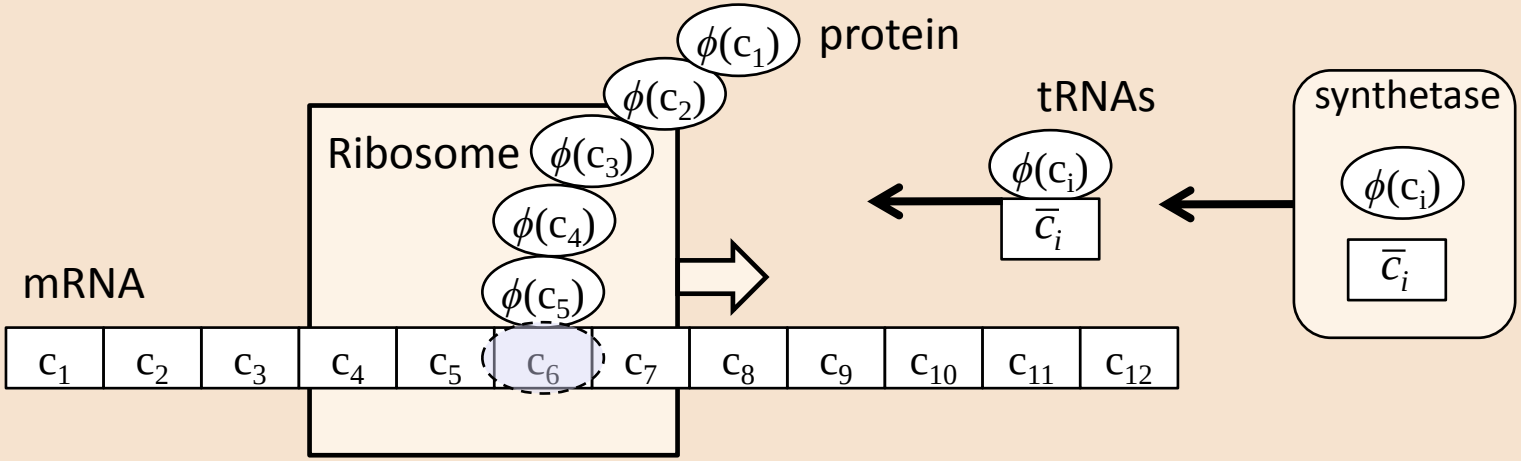
Fig. 2. Pancreatic exocrine cell. Array of cisternae of the rough surfaced endoplasmic reticulum.

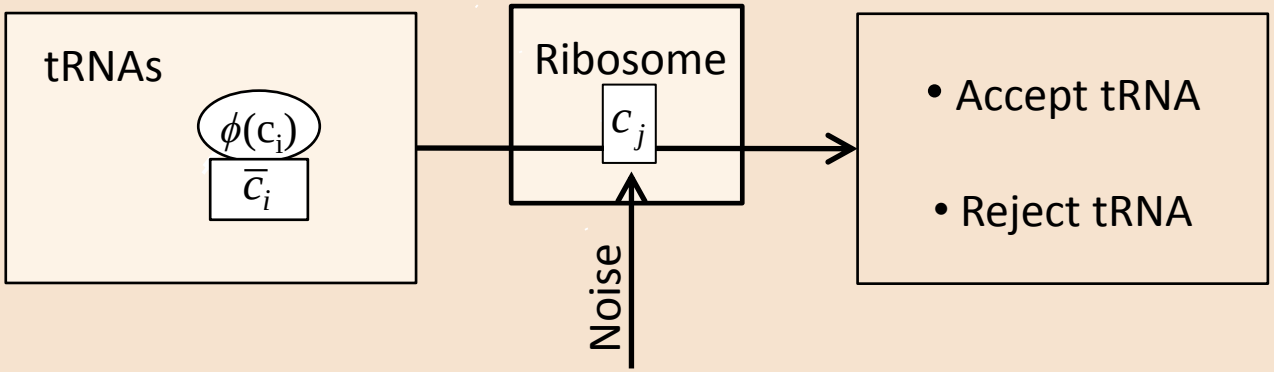
cs, cisternal space; cm, cytoplasmic matrix (cell sol); *fr*, free ribosomes; *ar*, attached ribosomes; *mer*, membrane of the endoplasmic reticulum.

x 50,000

George Palade (50s)

Ribosome needs to recognize the correct tRNA



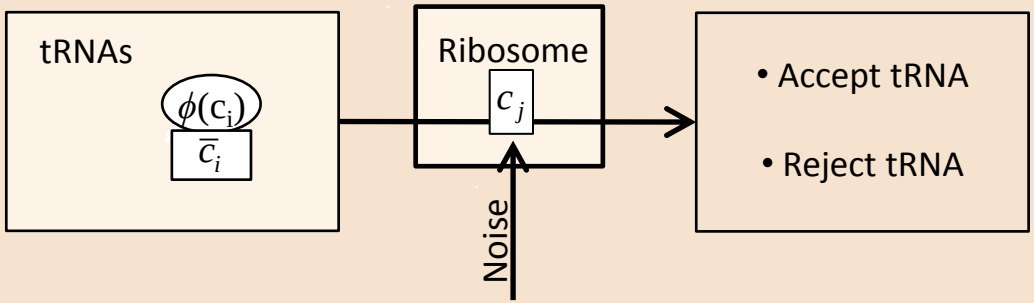


(i) binding wrong tRNAs: $\text{amino-acid} \neq \phi(\text{codon})$

(ii) unbinding correct tRNAs: $\text{amino-acid} = \phi(\text{codon})$

How to construct fast\accurate\small ***molecular*** decoder ?

2.Decoding at the ribosome is a molecular recognition problem



- Central problem in biology and chemistry:

How to evolve molecules that recognize in a noisy environment?

(crowded, thermally fluctuating, weak interactions).

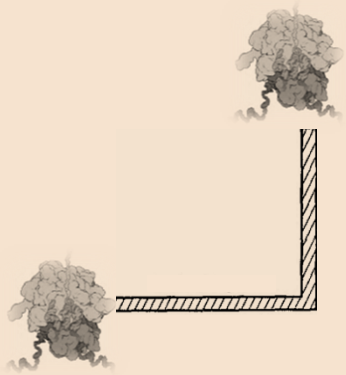
- How to estimate recognition performance ("fitness")?
- What are the relevant degrees-of-freedom? **Dimension**? **Scaling**?
- What is the role of conformational changes?

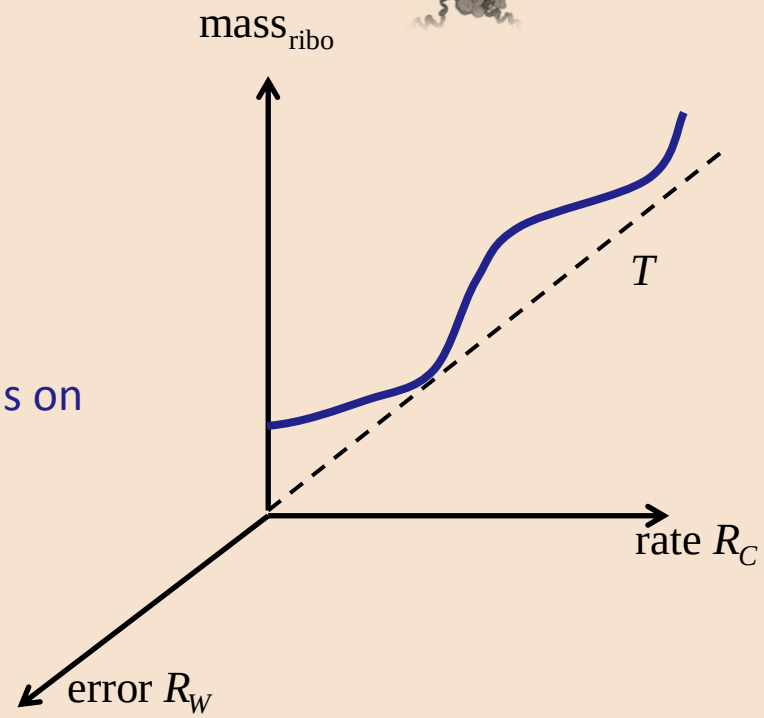
Ribosome sets physical limit on self-reproduction rate

Large fraction of cell mass is ribosomes.

- In self-reproduction each ribosome should self-reproduce.
- Sets lower bound on self-reproduction rate .

$$T \geq \frac{\text{mass}_{\text{ribo}}}{R_C} \approx \frac{10^4 \text{ amino-acids}}{20 \text{ amino-acids/sec}} = 500 \text{ sec}$$





Problem: how ribosome accuracy affects fitness depends on

(i) Basic protein properties (mutations).

(ii) Biological context (environment etc.).

Ribosomes are complicated machines with many d.o.f.

Ribosomes are made of proteins and RNAs:

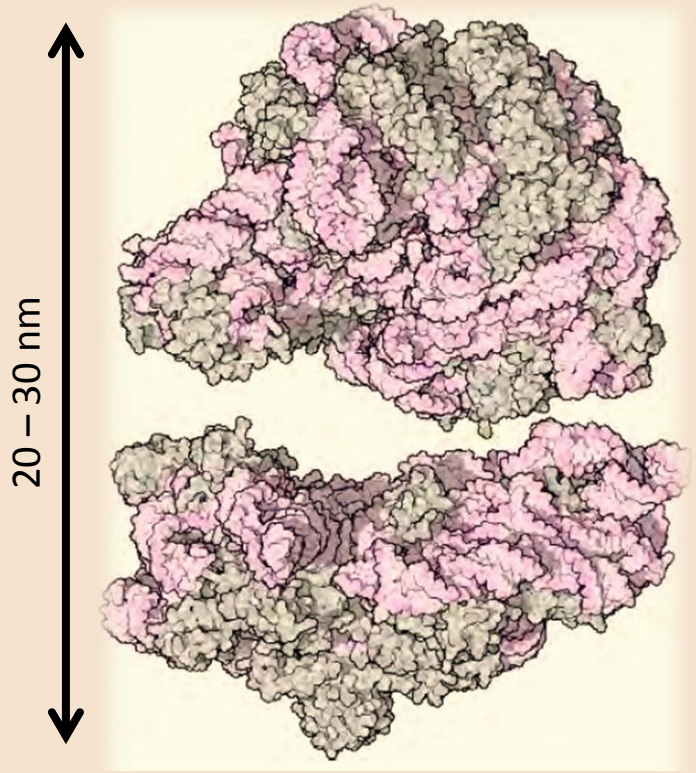
- ~ 10^4 nucleic bases in RNA.
- ~ 10^4 amino-acids in proteins.
- Total mass : ~ $3 \cdot 10^6$ a.u.

- High-res structure is known (Yonath et al.).

Within this known complexity:

- **What are the relevant degrees-of-freedom?**

- **How does this machine operate?**

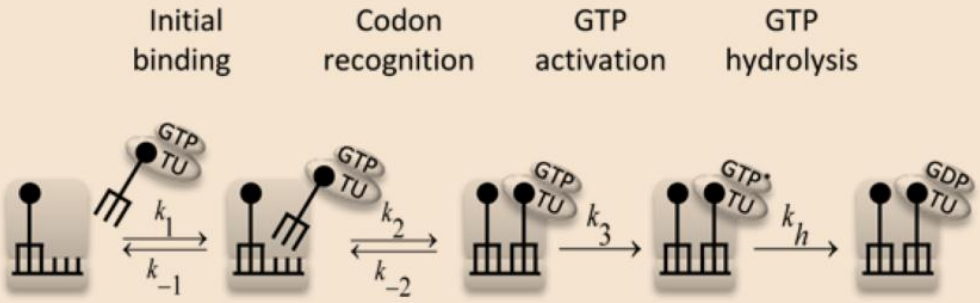


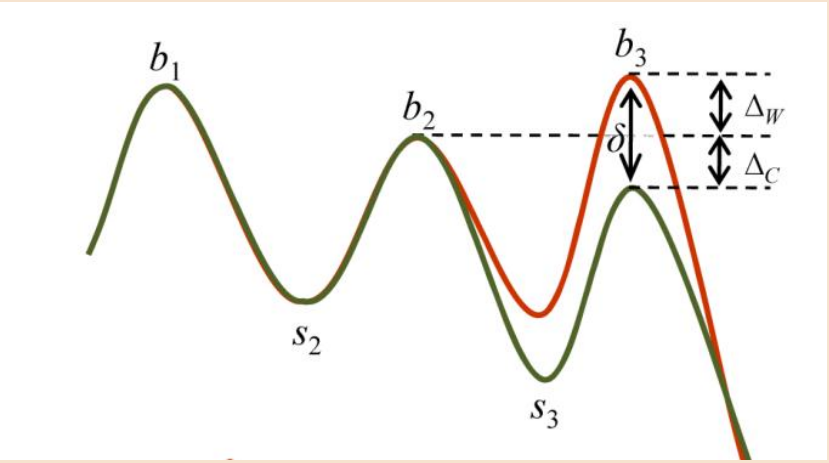
(magenta – RNA, grey – protein,

from Goodsell, *Nanotechnology*)

Decoding is determined by energy landscapes of correct and wrong tRNAs

- Decoding is multi-stage process.
- Kinetics involves *large* conformational changes.





Steady-state decoding rates

(Arrhenius law, $k \propto e^{-\Delta G}$)

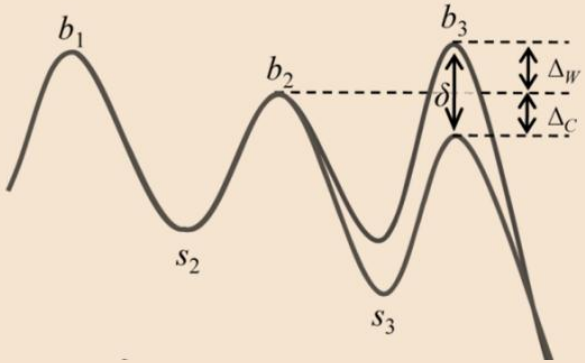
$$R_C \sim \frac{1}{e^{b_1} + e^{b_2} + e^{b_3}}$$

$$R_W \sim \frac{1}{e^{b_1} + e^{b_2} + e^{\boldsymbol{\delta}+b_3}}$$

In Ehrenberg's notation

- Merge the first two barriers to get Michaelis-Menten kinetics: $e^B = e^{b_1} + e^{b_2}$.

$$E + S^{c(\mathrm{nc})} \underset{k_{\mathrm{d}}^{c(\mathrm{nc})}}{\overset{k_a^{c(\mathrm{nc})}}{\rightleftarrows}} ES^{c(\mathrm{nc})} \xrightarrow{k_c^{c(\mathrm{nc})}} E + P^{c(\mathrm{nc})}$$



$$\frac{j^c}{s^c e} = \left(\frac{k_{cat}}{K_m}\right)^c = k_a^c \frac{k_c^c}{k_c^c + k_d^c} = k_a^c \frac{1}{1+a} \quad \Leftrightarrow \quad R_C = k_a^C \frac{1}{1+e^{b_3-B}} \propto \frac{1}{e^B + e^{b_3}}$$

$$\frac{j^{nc}}{s^{nc} e} = \left(\frac{k_{cat}}{K_m}\right)^{nc} = k_a^{nc} \frac{k_c^{nc}}{k_c^{nc} + k_d^{nc}} = k_a^{nc} \frac{1}{1+d_d a} \quad \Leftrightarrow \quad R_W = k_a^W \frac{1}{1+e^{b_3-B+\delta}} \propto \frac{1}{e^B + e^{b_3+\delta}}$$

$$A = \left(\frac{k_{cat}}{K_m}\right)^c \Big/ \left(\frac{k_{cat}}{K_m}\right)^{nc} = d_a \frac{1+d_d a}{1+a} = \frac{d_a + da}{1+a} \quad \Leftrightarrow \quad \frac{R_C}{R_W} = \frac{e^B + e^{b_3+\delta}}{e^B + e^{b_3}} = \frac{1+e^{b_3-B+\delta}}{1+e^{b_3-B}}$$

$$a = \frac{k_d^c}{k_c^c} = e^{b_3-B}; \qquad d_a = \frac{k_a^c}{k_a^{nc}} = 1;$$

$$d_d = \frac{k_d^{nc}}{k_c^{nc}} \Big/ \frac{k_d^c}{k_c^c} = e^{\delta}; \quad d = d_d d_a = e^{\delta};$$

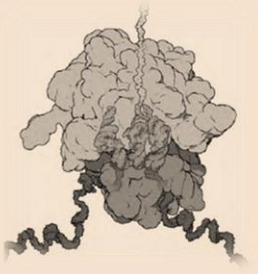
$$\left(\frac{k_{cat}}{K_m}\right)^c = k_a^c \frac{1}{1+a} = k_a^c \frac{d-A}{d-1}$$

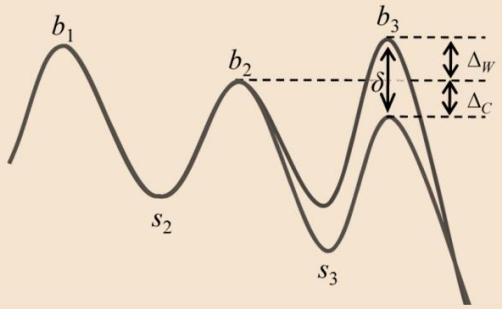
$$\Leftrightarrow \quad R_C = k_a^C \frac{1}{1+e^{b_3-B}} = k_a^C \cdot \frac{e^{\delta} - (R_C / R_W)}{e^{\delta} - 1}$$

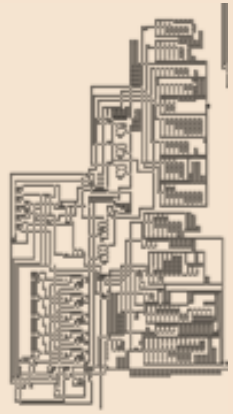
Ribosome kinetics exhibits large dimensionality reduction

- Effective dimension decreases by at least 3 orders of magnitude:

~ 10^4 structural parameters → ~ 10 kinetic parameters (energy landscape).







- Generic phenomenon in biomolecules: many catalytic molecules (enzymes) can be described by a few kinetic parameters (transition state landscape).

What is the origin of dimensionality reduction?

- Hints:

- Protein function mainly involve the lowest modes of their vibrational spectra (hinges).

- Sectors: “Normal modes” of sequence evolution (Leibler & Ranganthan).

Transition states reduce the dimensionality of effective parameter space

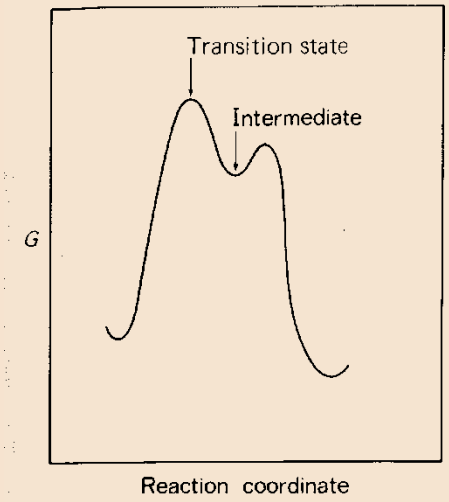


Figure 2.1 Transition states occur at the peaks of the energy profile of a reaction, and intermediates occupy the troughs.

$$k_1 = \left(\frac{kT}{h}\right)\exp\left(\frac{-\Delta G^{\ddagger}}{RT}\right)$$

FEBRUARY, 1935 JOURNAL OF CHEMICAL PHYSICS VOLUME 3

The Activated Complex in Chemical Reactions

HENRY EYRING, *Frick Chemical Laboratory, Princeton University*

(Received November 8, 1934)

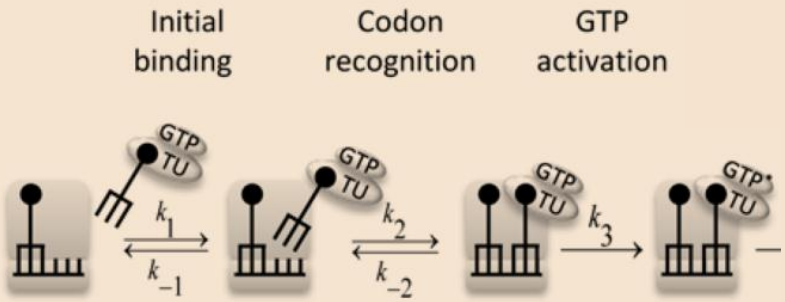
leave no doubt of the proposed method of procedure in a particular case. We may write for the specific reaction rate constant for a reaction of any order

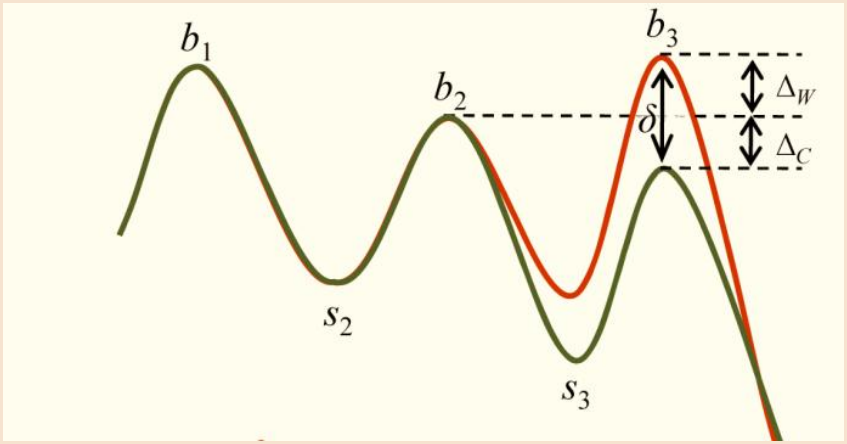
$$k_i = c(F_a/F_n)(\bar{p}/m^*) = c(F_a'/F_n)(kT/h)e^{-E_0/kT} \qquad (10)$$

where F_a is simply the partition function (or Zustandsumme) for the activated state and F_n is the same quantity for the normal state. F_a' is the partition function for the activated complex for all the normal coordinates except the one in which decomposition is occurring. The partition function for this normal coordinate is included in the factor $(kT/h)e^{-E_0/kT}$. The other quantities have been defined.

Theory can be tested with measured rates

- The codon-specific stages are Codon recognition and GTP activation.





	(UUU) Cognate	(CUC) Non-cognate
k_1	100-140 1/(μM·s)	100-140 1/(μM·s)
k_{-1}	80-100 1/s	80-100 1/s
k_2	190 1/s	190 1/s
k_{-2}	0.23 1/s	100 1/s
k_3	260 1/s	0.6 1/s

(Rodnina's lab, Gottingen)

$$R_C \sim \frac{1}{e^{b_1} + e^{b_2} + e^{b_3}}$$

$$R_W \sim \frac{1}{e^{b_1} + e^{b_2} + e^{\boldsymbol{\delta}+b_3}}$$

How to estimate recognition performance (“fitness”) ?

What is the actual dimension of the problem ?

Recognition fitness has generic features

$$R_C \sim \frac{1}{e^B + e^{b_3}}$$

$$R_W \sim \frac{1}{e^B + e^{\delta + b_3}}$$

$$(e^B = e^{b_1} + e^{b_2})$$

- "Fitness" *F* is often obscure and context-dependent:

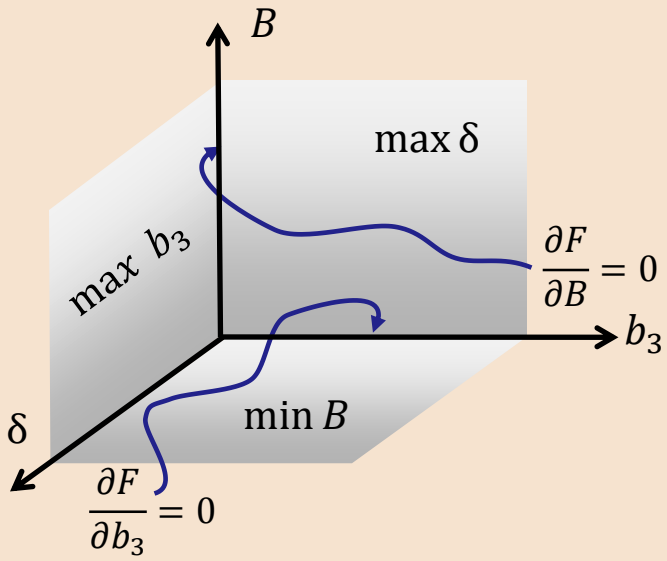
  → look for generic properties of $F(R_C, R_W) = F(B, \delta, b_3)$.

- Only requirement: "biologically reasonable", $\frac{\partial F}{\partial R_C} \geq 0$, $\frac{\partial F}{\partial R_W} \leq 0$.

- Searching for optimum in (B, δ, b_3) space:

(i) $\frac{\partial F}{\partial \delta} \geq 0$: $\boldsymbol{\delta}$ **approaches biophysical limit**.

(ii) $\left\{\frac{\partial F}{\partial B} = 0 \;\&\; \frac{\partial F}{\partial b_3} \geq 0\right\}$ or $\left\{\frac{\partial F}{\partial B} \leq 0 \;\&\; \frac{\partial F}{\partial b_3} = 0\right\}$:



Optimization is essentially 1D (2 other parameters approach limit).

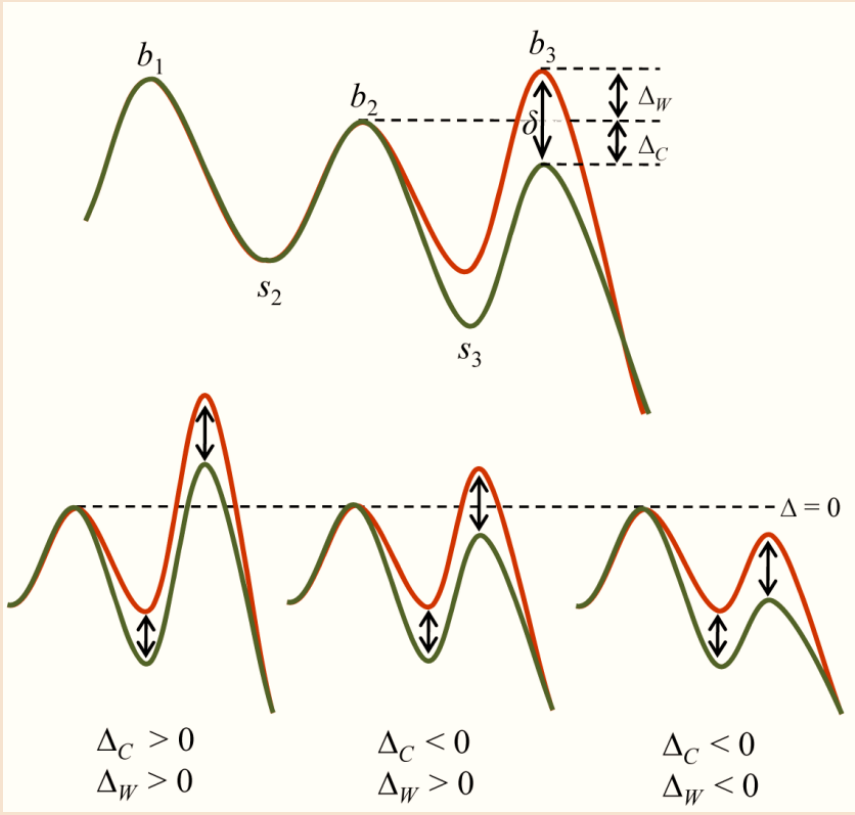
What is the optimal energy landscape of the ribosome?

- For example, distortion fitness from engineering (weight d is context-dependent) :

$$F = R_C - d \cdot R_W \propto \frac{1}{e^B + e^{b_3}} - \frac{d}{e^B + e^{b_3+\delta}}$$

- 1D problem: optimum is along b_3.

(measured: $\Delta = b_3 - b_2$)



- **What is the optimal b_3 (or Δ)?**
- **Is the ribosome optimal ?**
- **Role of conformational changes ?**

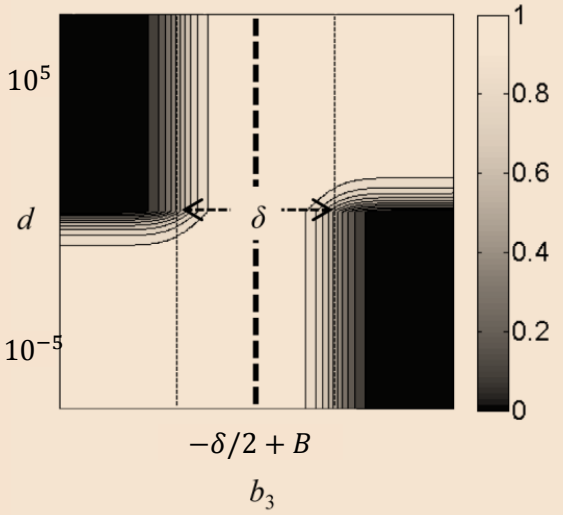
Optimal design is a Max-Min strategy

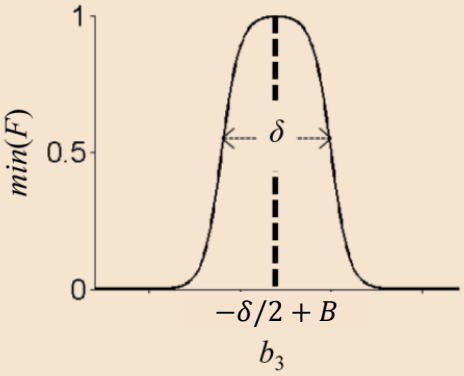
- Weight d can vary.

(i) For each d normalize F.

(ii) "Worst case scenario":

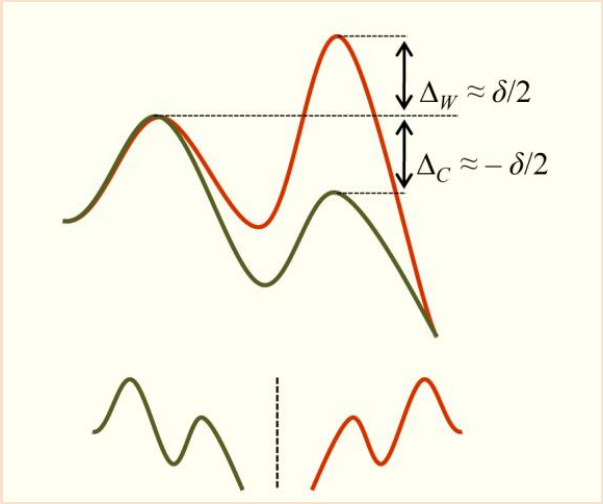
$$\max\big(\min(F)\big).$$





- Max-Min solution is "symmetric":

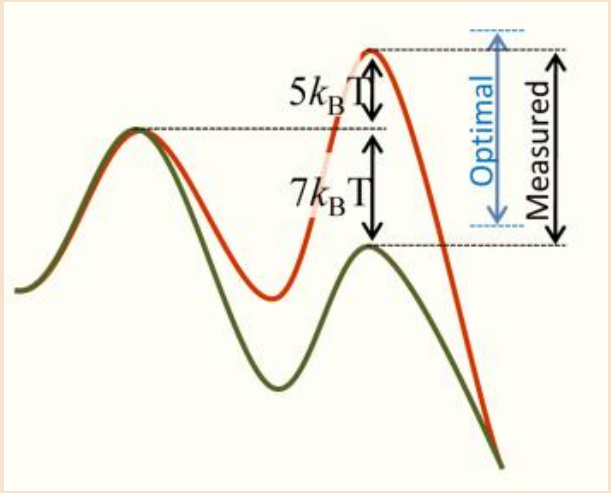
$$b_3 = -\frac{1}{2}\delta + B$$

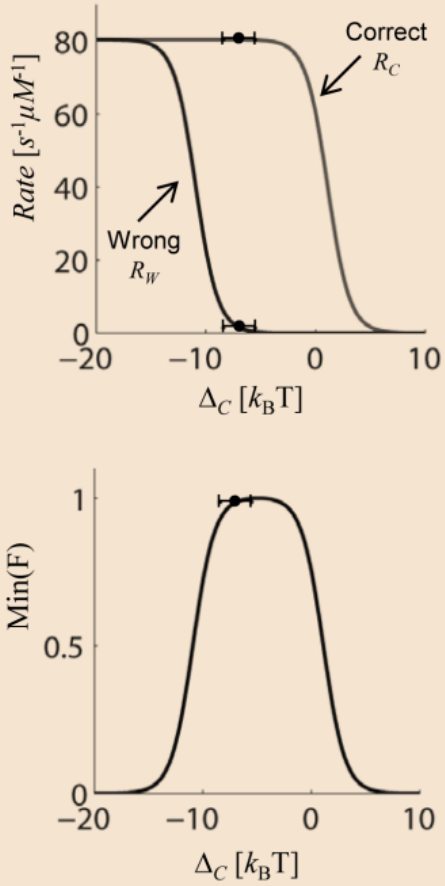


Ribosome shows an energy barrier which is nearly optimal

- Measurements: $\Delta_C \approx -7\ k_B T,\ \delta \approx 12 k_B T\ , \bar{B} = B - b_2 \approx 1 k_B T.$
- Prediction: the optimal regime is **symmetric**, $\Delta_C < 0,\ \Delta_W > 0.$

$$\Delta_C = -\frac{1}{2}\delta + \bar{B}$$

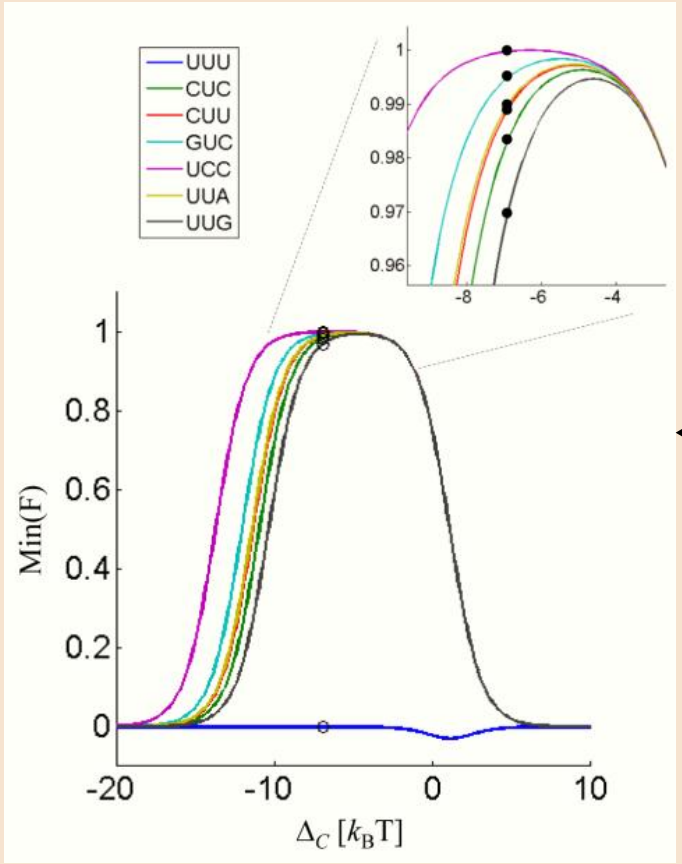


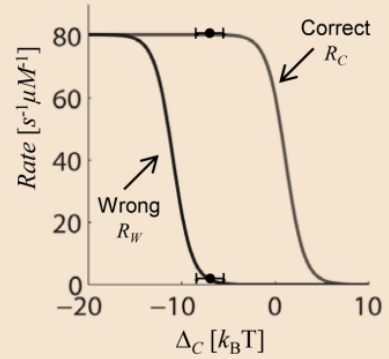


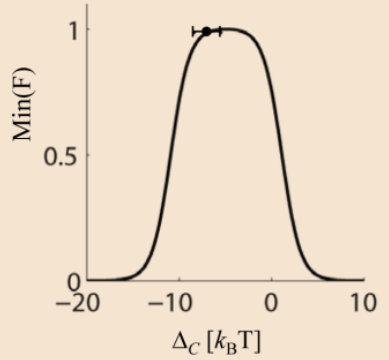
- **The ribosome is nearly optimal**

  (according to Max-Min prediction).

Decoding is optimal for all six measured tRNAs







- Except for UUC which encodes the same amino-acid

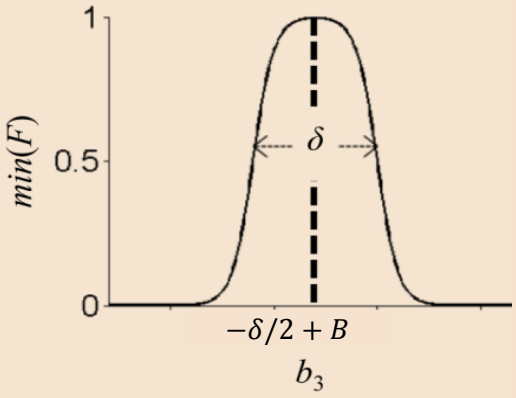
$$\phi(\mathrm{UUC}) = \phi(\mathrm{UUU}) = \text{phenylalanine.}$$

Optimality is valid for wide range of fitness functions

- Ribosome optimal in wide region:

$$5\cdot 10^{-6} = e^{-\delta} \le d \le e^{\delta} = 2\cdot 10^{5}.$$

$F = R_C - d \cdot R_W$

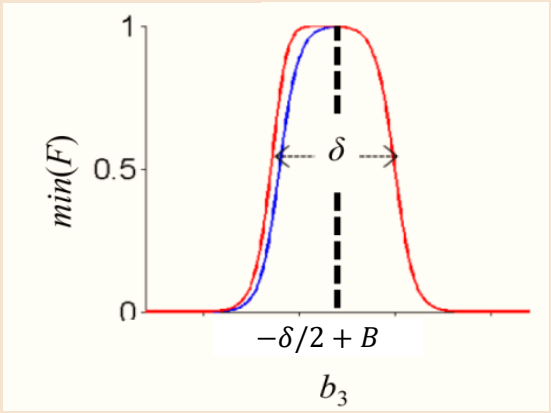


- **General feature**: any fitness function $F(R_C, R_W)$

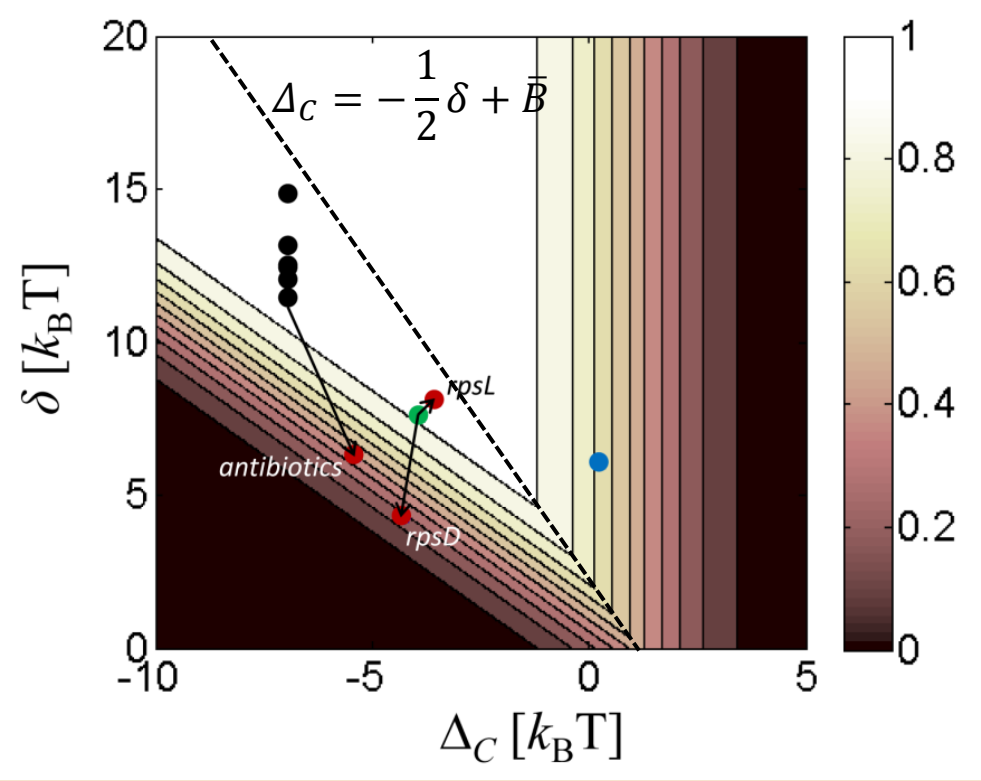
exhibits optimum as long as both rates are "relevant".

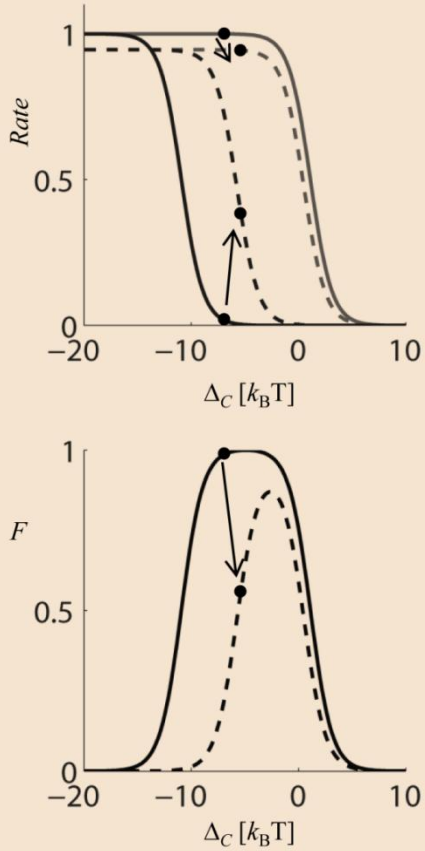
$$e^{-\delta} < \left| \frac{\partial F}{\partial R_C} \Big/ \frac{\partial F}{\partial R_W} \right| < e^{\delta}$$

$R_C - d \cdot R_W{}^2$ $R_C - d \cdot (R_W/R_C)$



Theory predicts optimal regime of ribosomes for all organisms



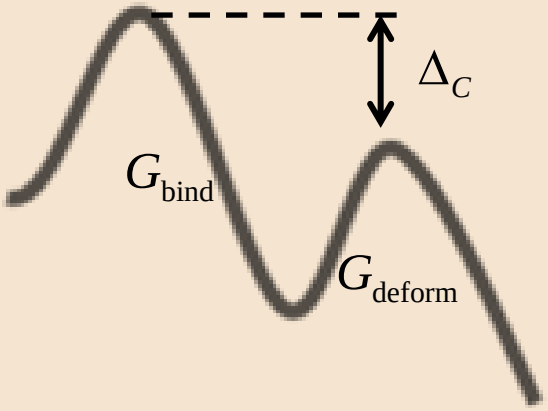


- Optimal region in the space of all possible landscapes, $-\delta \leq \Delta_C \leq 0$.
- Mutations and antibiotics tend to push away of optimality.

What is the role of conformational changes?

- Energy barrier results from binding energy and deformation energy penalty:

$$\Delta_C = G_{\text{deform}} - G_{\text{bind}} = -\frac{1}{2}\delta + \overline{B}.$$



- Therefore

$$G_{\text{deform}} = G_{\text{bind}} - \frac{1}{2}\delta + \overline{B}.$$

- For any $G_{\text{bind}} \geq -\frac{1}{2}\delta + \overline{B} \approx 5\ \text{k}_\text{B}\text{T},$

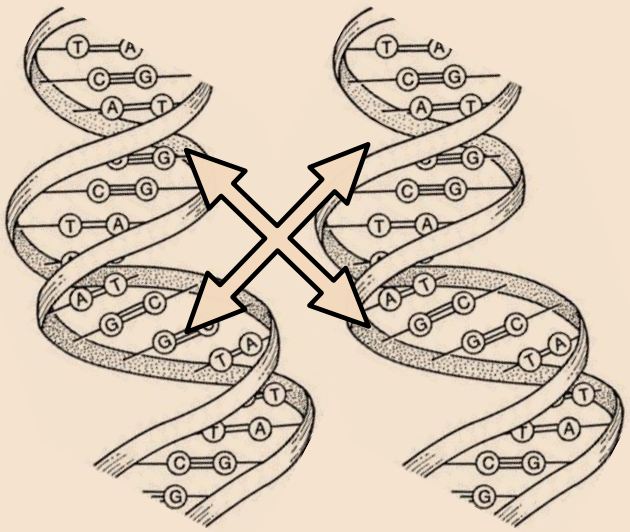
$G_{\text{deform}} \geq 0$ non-zero deformation is optimal for tRNA recognition.

Energy barrier that discerns the right target from competitors.

Yonatan Savir

Recombination machinery recognizes homologous DNA

- Exchange between two *homologous* DNAs.
- Essential for:
  - *Genome integrity* (repair machinery).
  - *Genetic diversity* (crossover, sex).

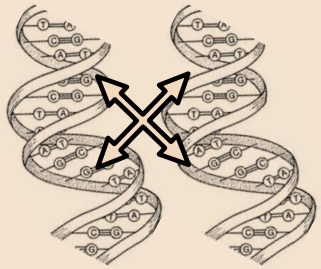


- Task: Detect correct, homologous DNA target among many incorrect lookalikes.

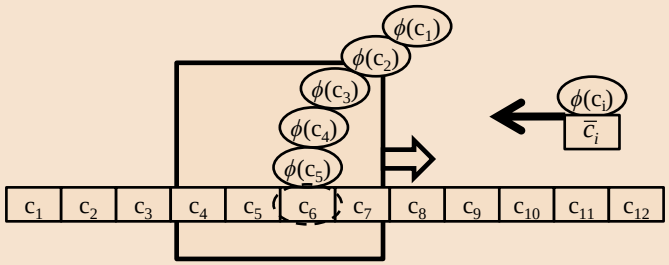
- DNA stretches during recombination:

  large deformation energy barrier.

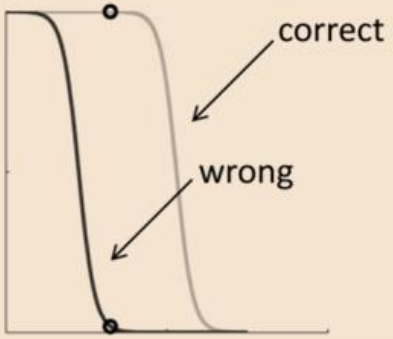
Energy barriers for optimal recognition may be a general design principle of recognition systems with competition



Recombination optimizes extension energy of dsDNA.



Ribosome optimizes energy barriers of decoding



Relevant energy

- **Conformational proofreading:** Design principle follows from optimization of information transfer function.
- May explain **induced fit** (Koshland 1958). Why molecules deform upon binding to target.

Relevant energy

- **Applies to any enzymatic kinetics in the presence of competition...**

Open questions, future directions…

Understanding evolvable matter:

- What are the degrees-of-freedom underlying dimensional reduction?
  (Rubisco and other enzymes)

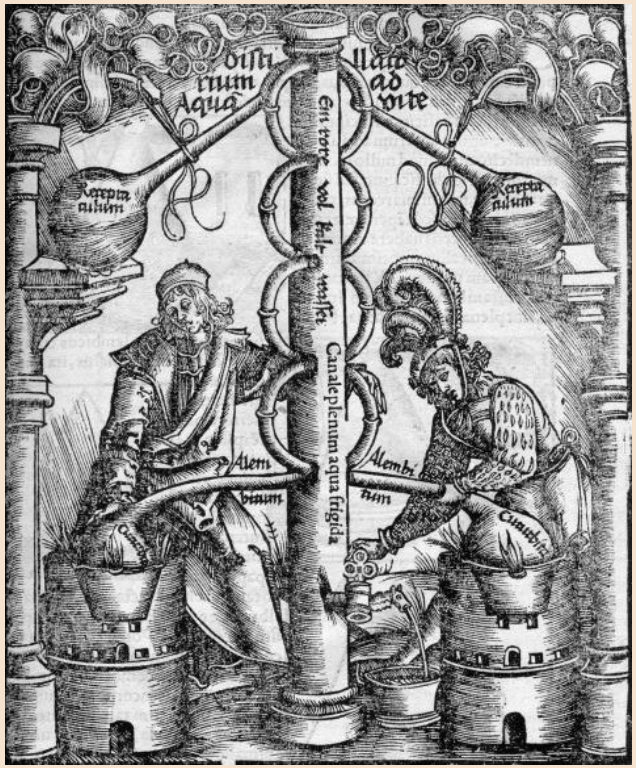
- Basic logic of molecular information channels
  (e.g. utilizing conformational changes, worst case scenario).

Translation machinery coevolved with proteins →

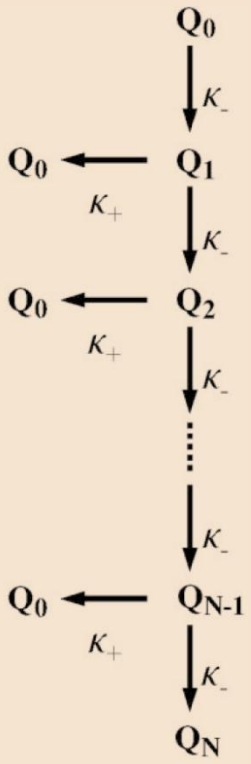
- Physics of the state of matter called “proteins”
  (evolvable, mapped from DNA space, glassy dynamics) .

Kinetic Proofreading

- The basic idea: iterations of irreversible discrimination step lead to exponential amplification.



Hopfield (1974), Ninio (1975)

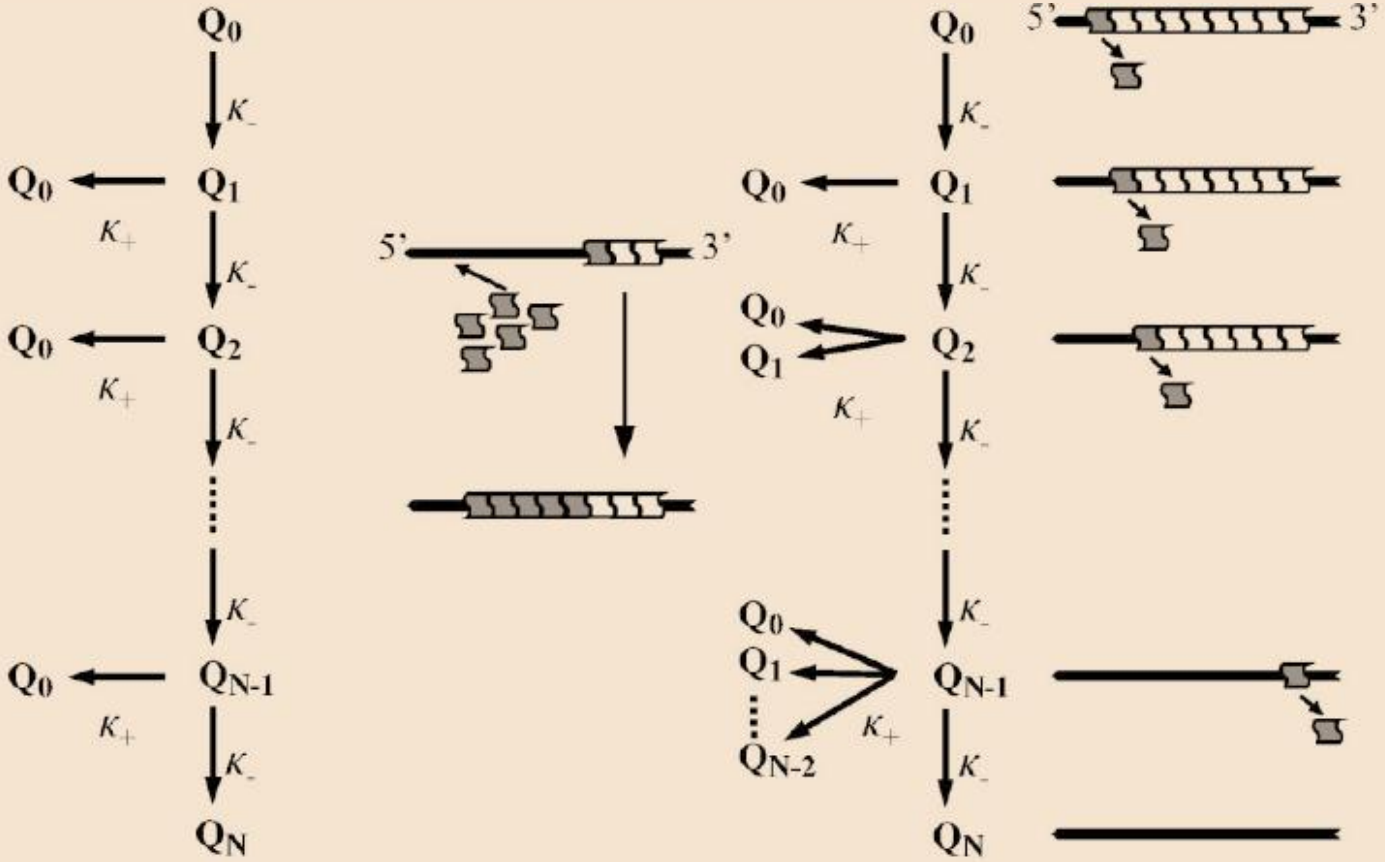


$$Q_N \sim K^{-N} = \left(\frac{\kappa_+ + \kappa_-}{\kappa_-} \right)^{-N}$$

$$F \equiv \frac{Q_N^C}{Q_N^W} = \left(\frac{K^C}{K^W} \right)^{-N} = \alpha^N$$

Bar-Ziv Libchaber (2002,2004)

RecA dynamics exhibits multistage KPR



$$F = \alpha^{N}$$

$$F = \alpha^{N^2/2}$$

RecA filament strongly fluctuates

- Gradual depolymerization vs. polymerization jumps (similar to microtubules):

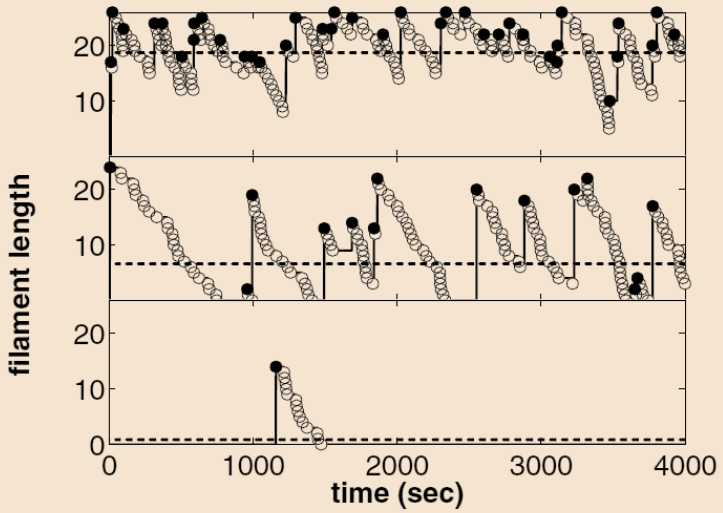
$$\frac{dp_n(t)}{dt} = -\kappa_-[p_n(t) - p_{n-1}(t)] + \kappa_+\left[\sum_{m=n+1}^{N} p_m(t) - np_n(t)\right]$$

- Continuum approximation $(p_n = P(n \text{ vacancies}),\ P_n = \sum_{m=n}^{N} p_m)$

$$\frac{\partial P(n,t)}{\partial t} = -\kappa_- \frac{\partial P(n,t)}{\partial n} - \kappa_+ nP(n,t)$$

- Fluctuations "scan" the sequence and

are therefore sensitive to mutations.



RecA dynamics is ultra-sensitive to DNA sequence

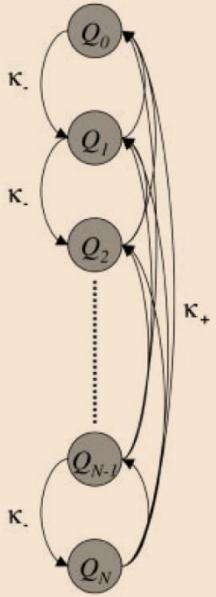
- At steady-state Gaussian amplification

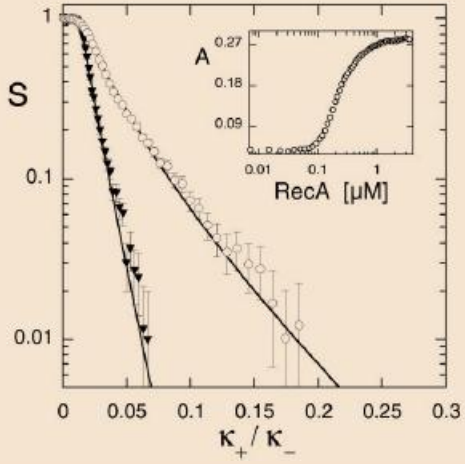
$$P_s(n) = e^{-(\kappa_+/2\kappa_-)n^2}$$

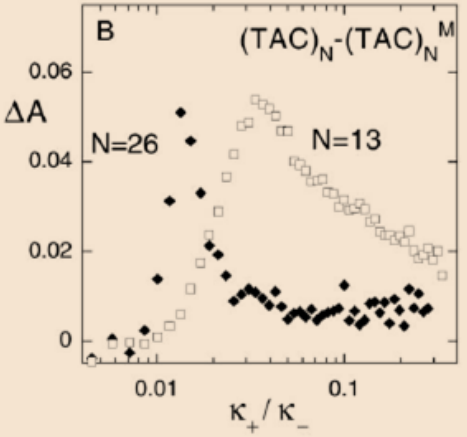
- General result (Murugan, Huse & Leibler):

$$P \sim \exp(-\#\ \text{loops}).$$

- Can sense even single mutations:





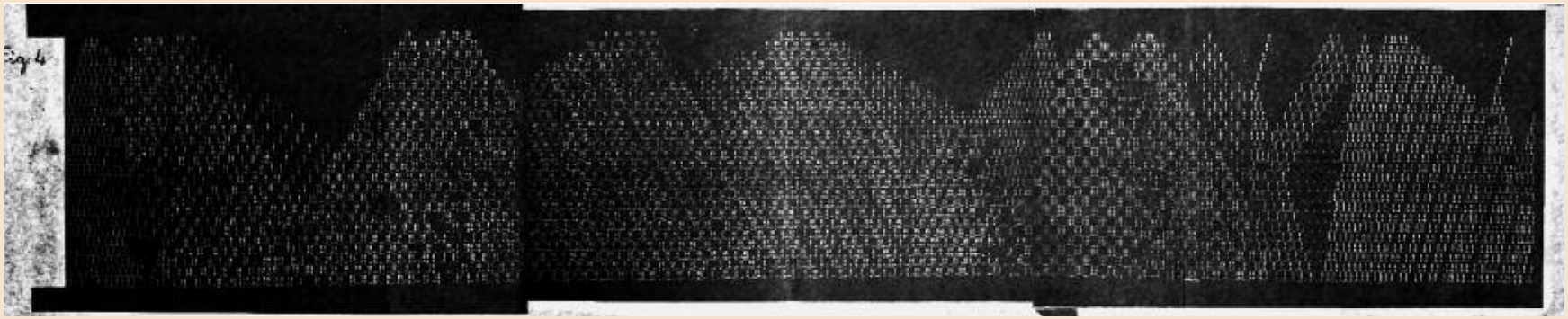


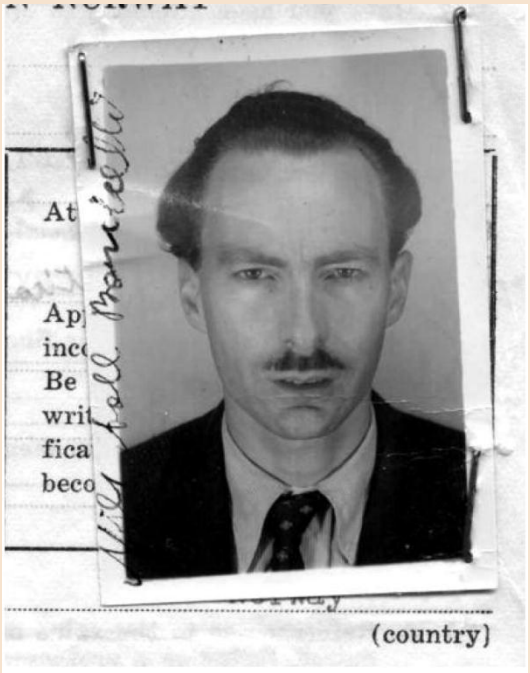
THANKS

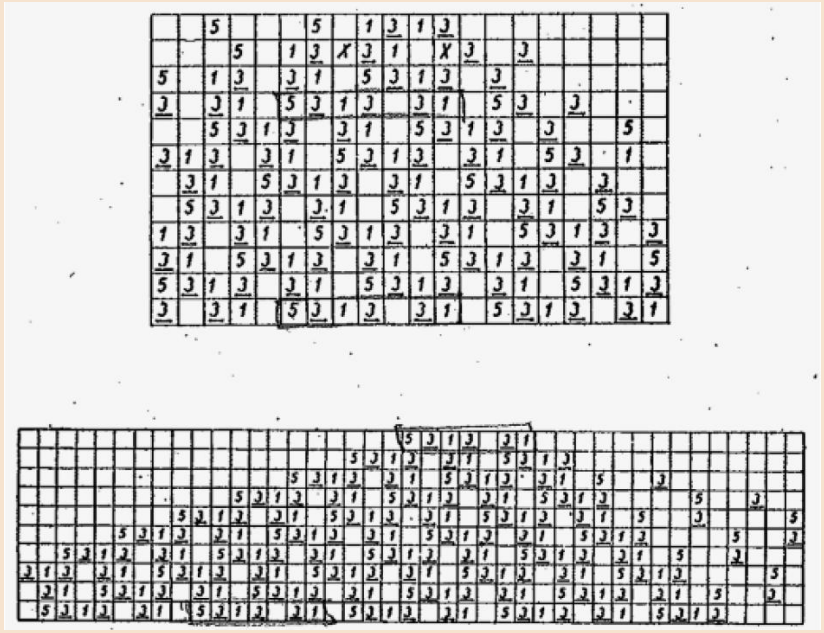
Yonatan Savir

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

Early simulations of artificial “evolution”







Niels Aall Barricelli