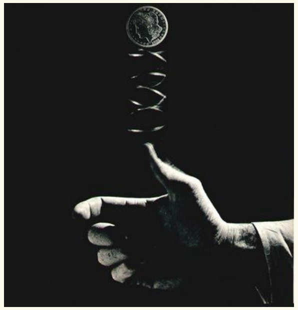
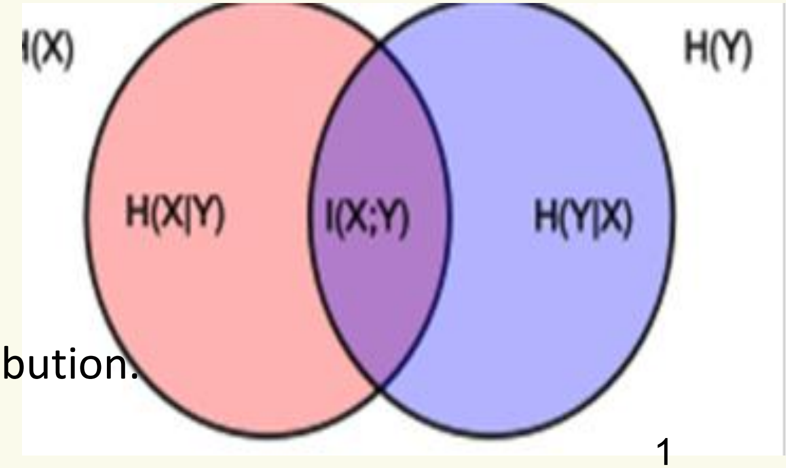
In previous lecture

- Shannon's information measure $$H(X) = -\sum_x p_x \log_2 p_x = -\langle \log_2 p_x \rangle$$



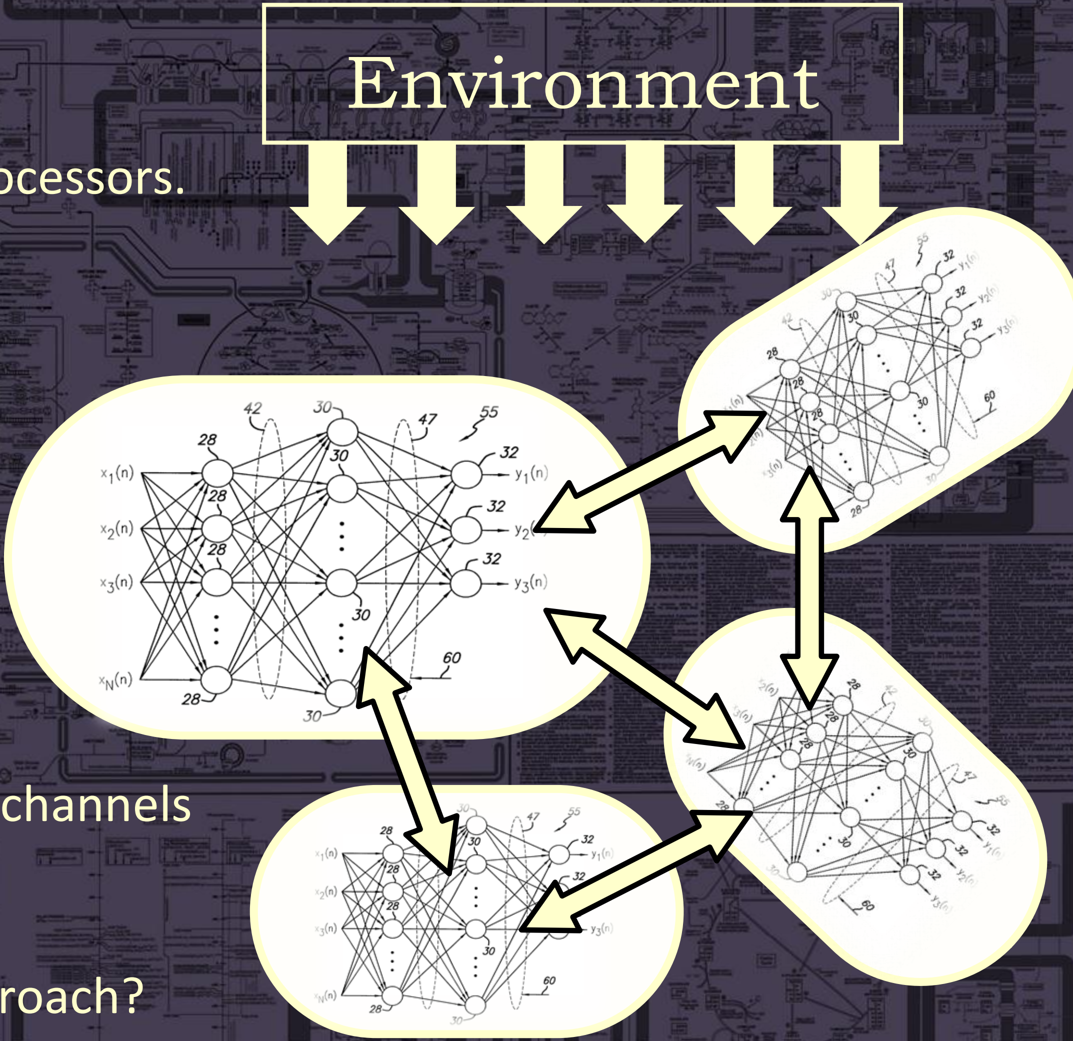
- Intuitive notion: *H* = number of required yes/no questions.
- The basic information unit is **bit =** 1 yes/no question or coin flip .
- For two variables X, Y we can measure how much they tell about each other.
- Information has *thermodynamic meaning:* One can produce mechanical work from information (Maxwell's demon).



- **Maximum Entropy** yields the most probable distribution.

II. Living Information – Overview: molecules, neurons, evolution, population.

Biological information is carried by molecular recognition



"Living systems"

I. Self-replicating information processors.

II. Evolve collectively.

III. Made of molecules.

- Generic properties of molecular channels subject to evolution?

- Physical/information theory approach?

Living Information - Overview: molecules, neurons, population and evolution

Living systems as information processors:

- *Sources of information in Life*: sequential information, cell composition, environment, population composition.

- *Living information channels*: (modernized) central dogma, replication, codes, receptors – signaling pathways, population dynamics, quorum sensing.

- *Information processing*: circuits and their elements, neural networks, genetic networks.

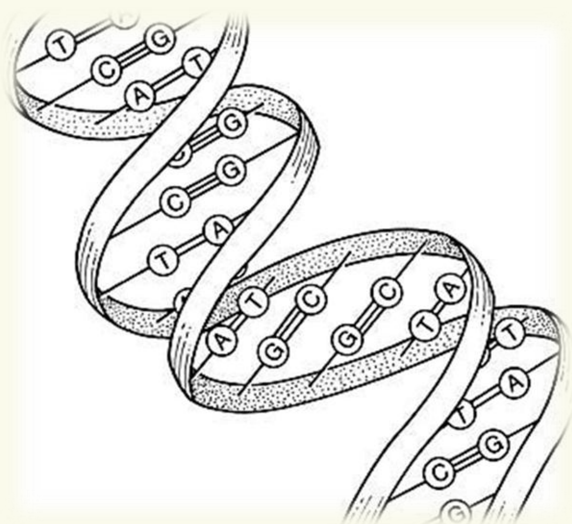
- *Information output*: transcription level, decisions, cell fate, differentiation and development; feedback (information loop).

Sequential information in DNA and proteins

DNA

- Building blocks :

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

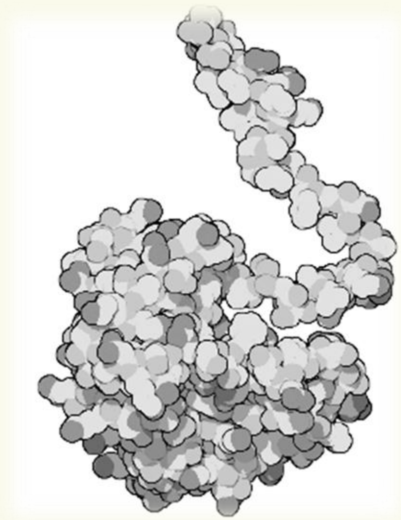


- Polymer: DNA double-helix.
- Inert information storage (“tape”)

protein

- Building blocks:

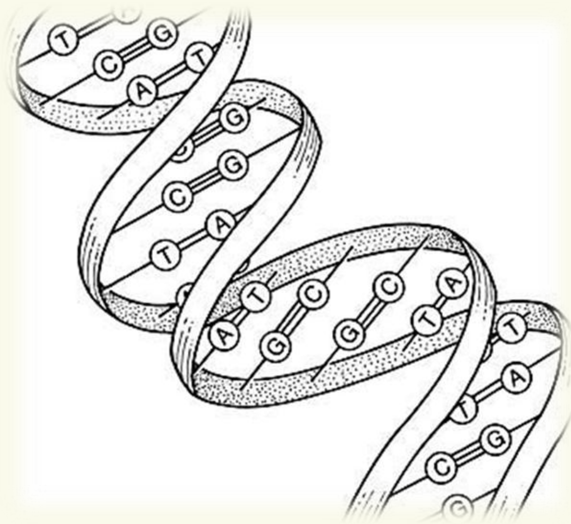
  20 amino acids.



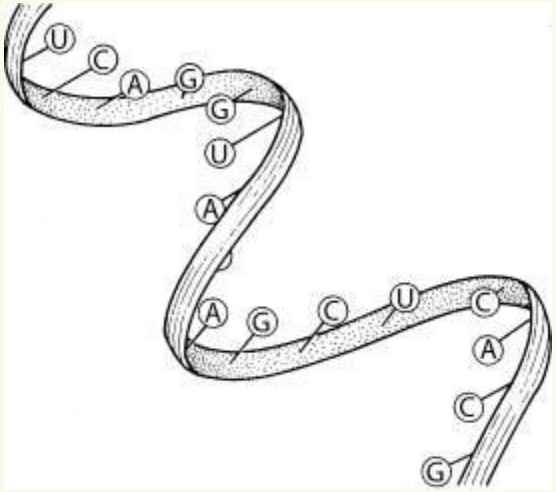
- Polymer = protein.
- Functional molecules (“constructor”)

RNA intermediates can be both tapes and machines

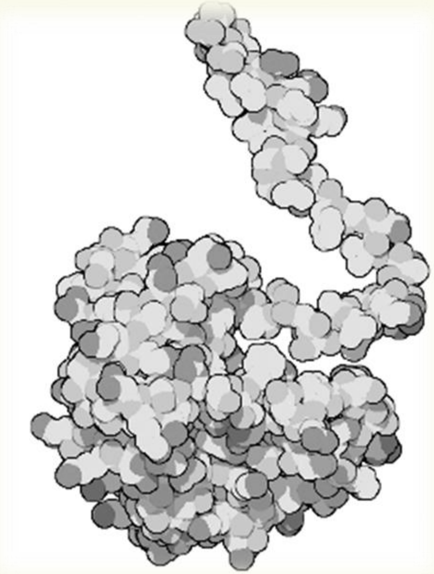
DNA



RNA



protein

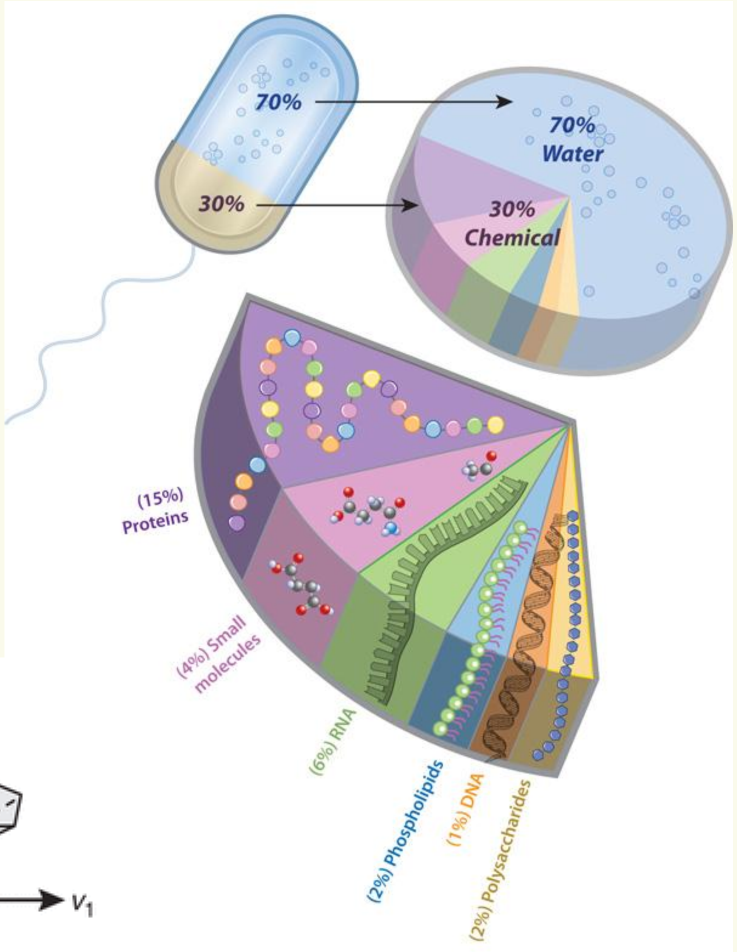


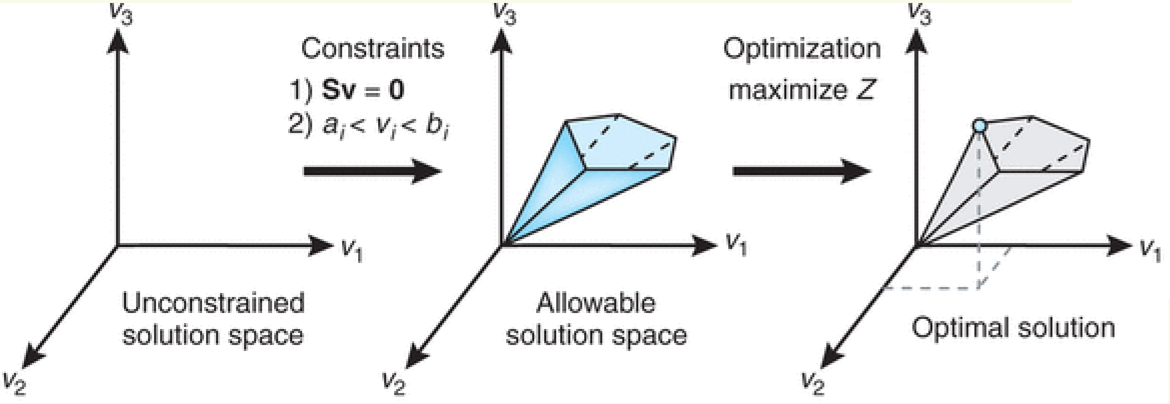
- Primordial “RNA world” :

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

Information in cellular composition

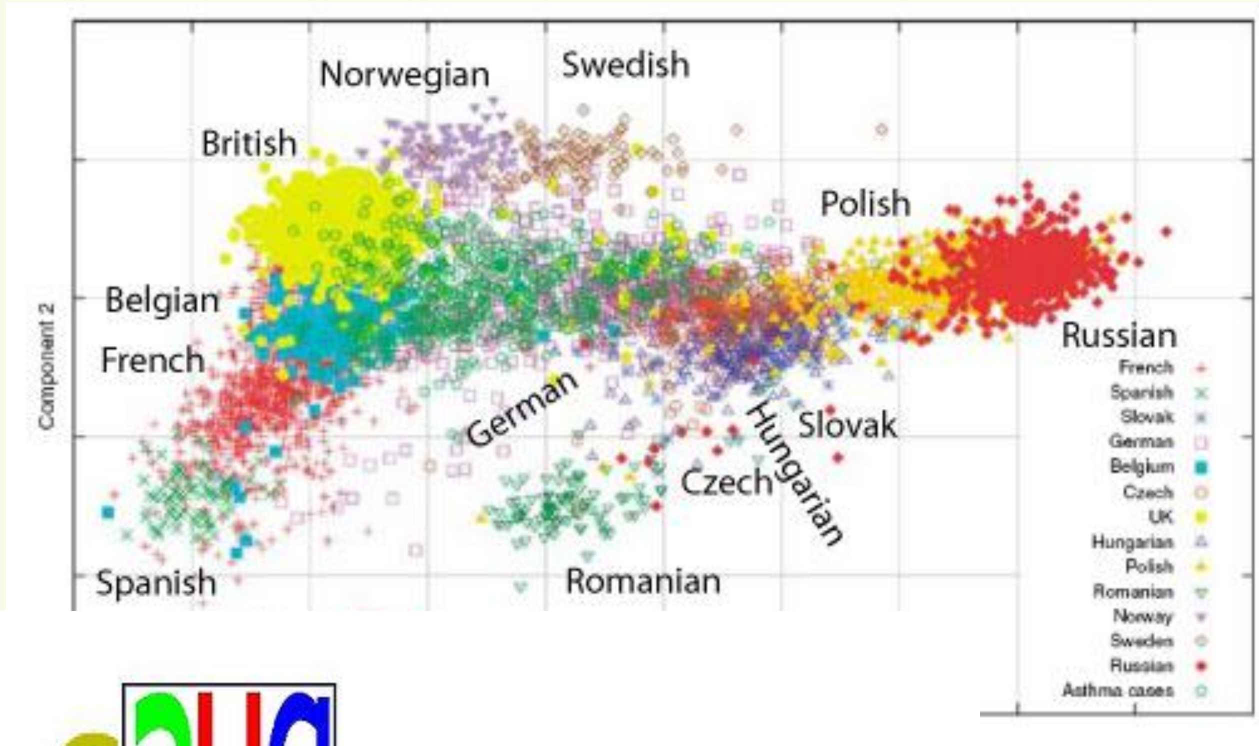
- Proteome, metabolome, lipome and other x-omes.
- Information can be represented as a composition vector.
- What is the relevant information?

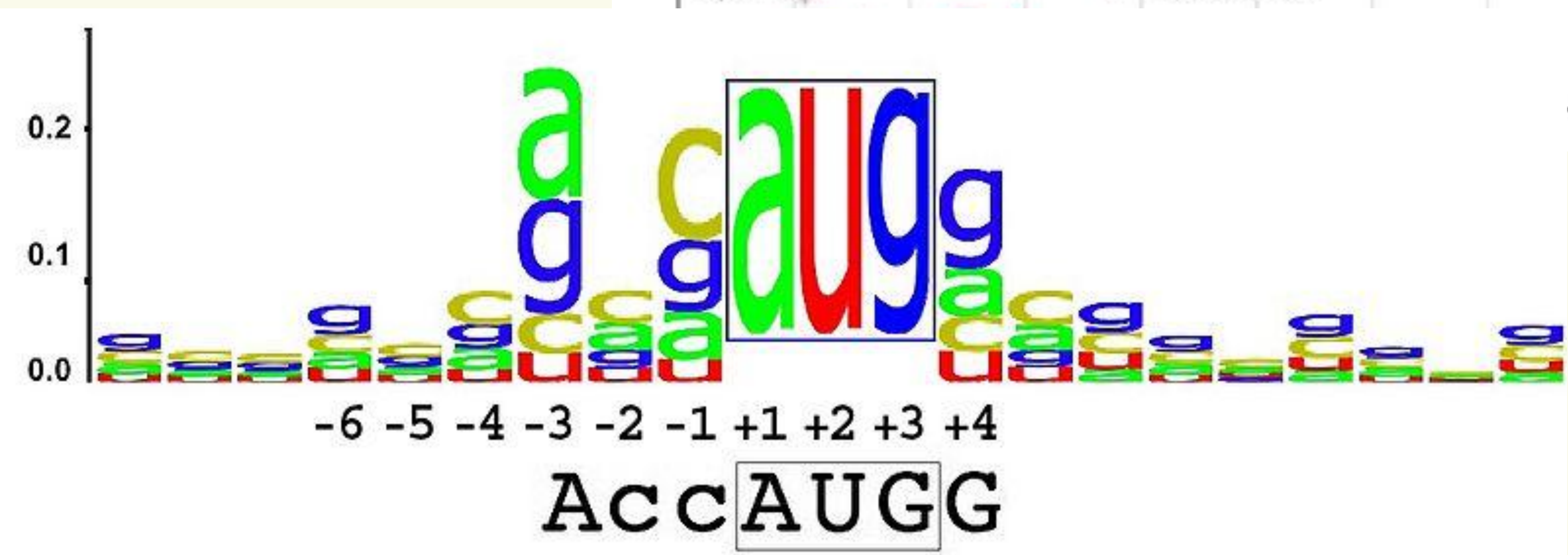




Information in population structure

$$R = 2 - H \quad \text{bit}$$





Living Information - Overview: molecules, neurons, population and evolution

Living systems as information processors:

- *Sources of information in Life*: sequential information, cell composition, environment, population composition.

- *Living information channels*: (modernized) central dogma, replication, codes, receptors – signaling pathways, population dynamics, quorum sensing.

- *Information processing*: circuits and their elements, neural networks, genetic networks.

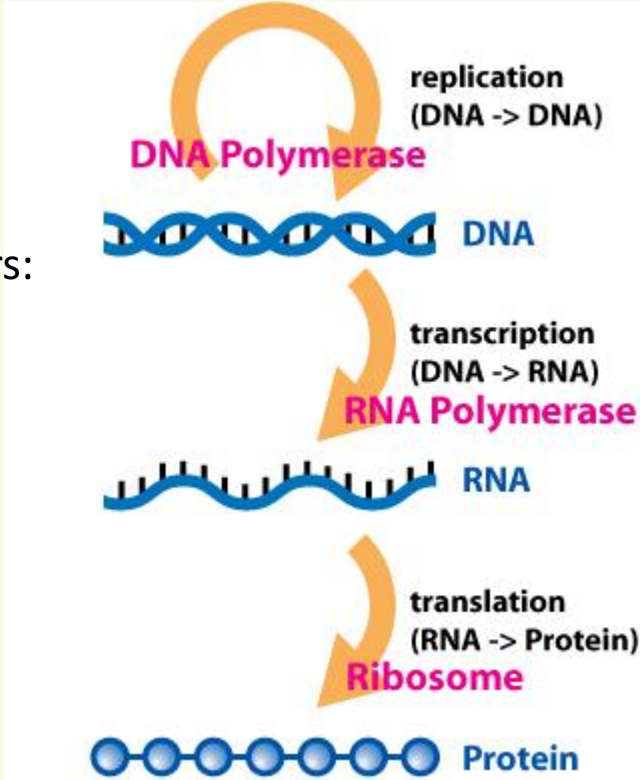
- *Information output*: transcription level, decisions, cell fate, differentiation and development; feedback (information loop).

Information in molecules: The central dogma

- **The central dogma of molecular biology** (Crick 1958, Nature 1970):

Information about DNA sequence cannot be transferred back from protein to either protein or nucleic acid.

- In M. Nirenberg words: *DNA makes RNA makes protein.*



- How sequence information is transferred between information-carrying biopolymers?
- 3 carriers: DNA, RNA protein and 3×3 = 9 potential transfers:

(i) 3 general transfers (occur in most cells).

(ii) 3 special transfers (only under specific conditions)

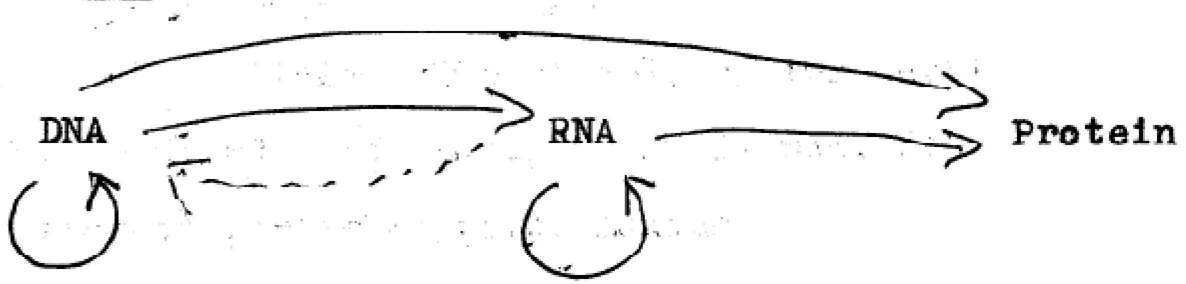
(iii) 3 unknown transfers (believed never to occur).

- General transfers: replication, transcription, translation.

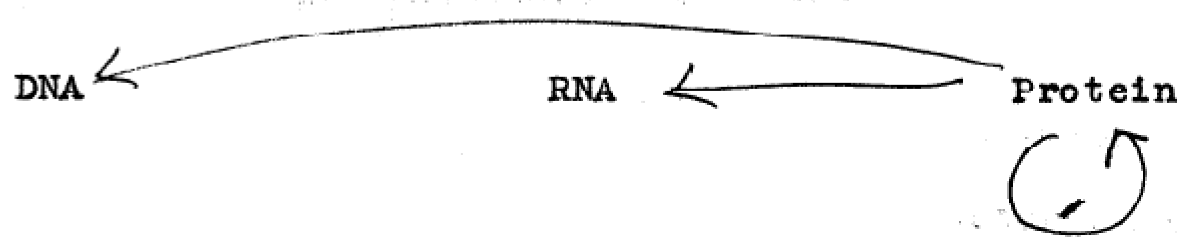
From Crick's draft (1956)

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.

“special” information transfers

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

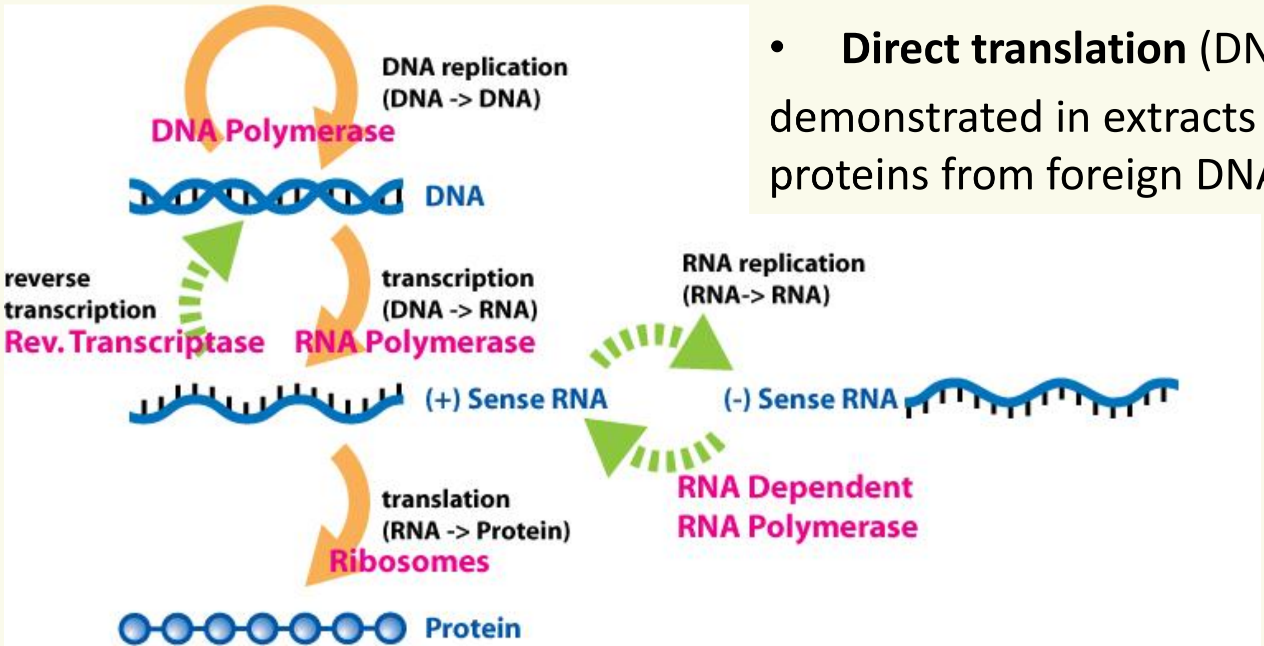
In retroviruses (HIV) and eukaryotes (retrotransposons and telomere synthesis).

- **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 DNA templates.



Transfers outside the central dogma

- **Posttranslational modification**

Protein amino acid sequence edited after translation by various enzymes.

- **Methylation**

Changes in methylation of DNA alter gene expression levels (usually DNA methylase). Heritable change is called **epigenetic .**

Effective information change but not primary DNA sequence

- **Prions**

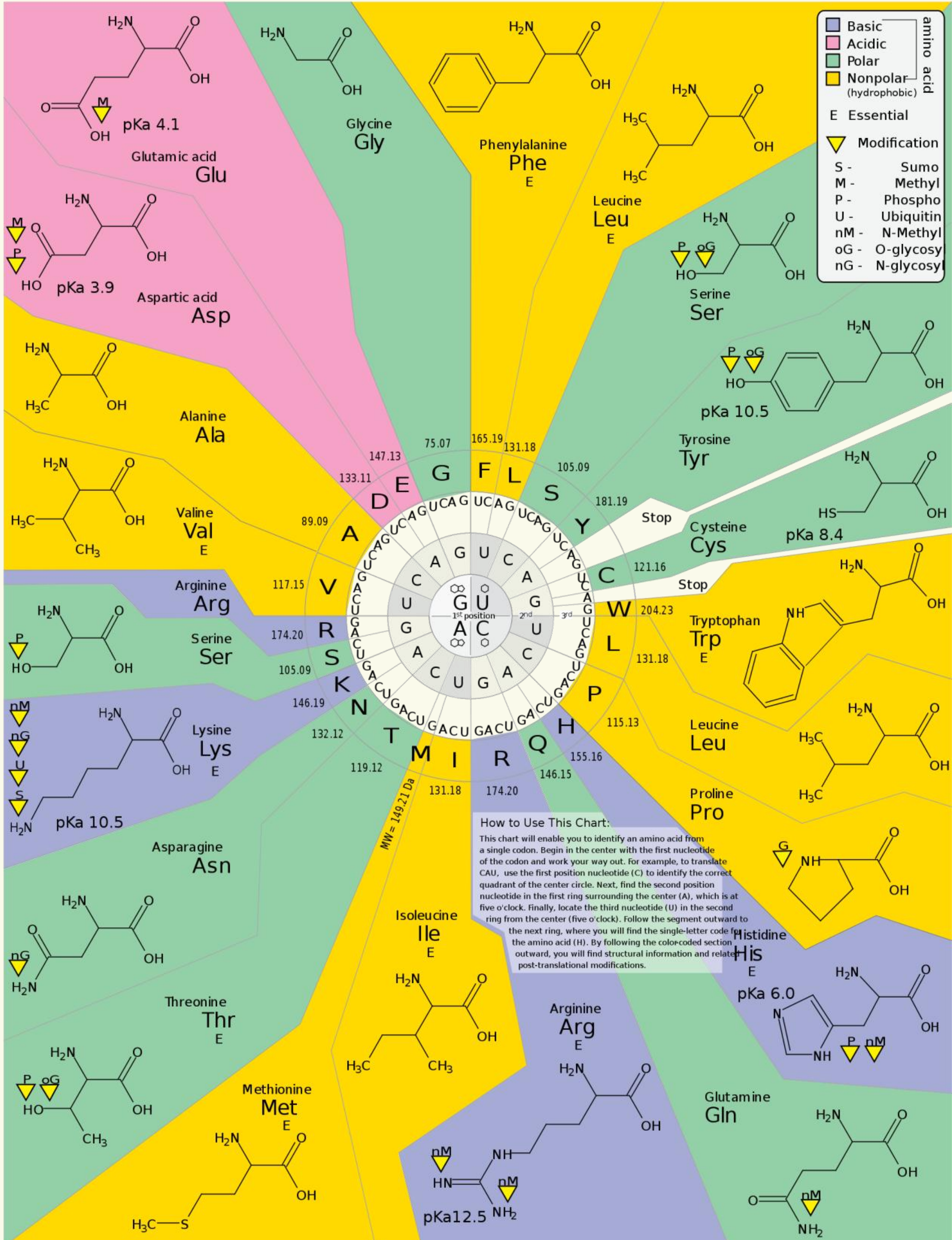
Proteins that propagate by making conformational changes in other molecules of the same protein.

Information propagated is protein conformation.

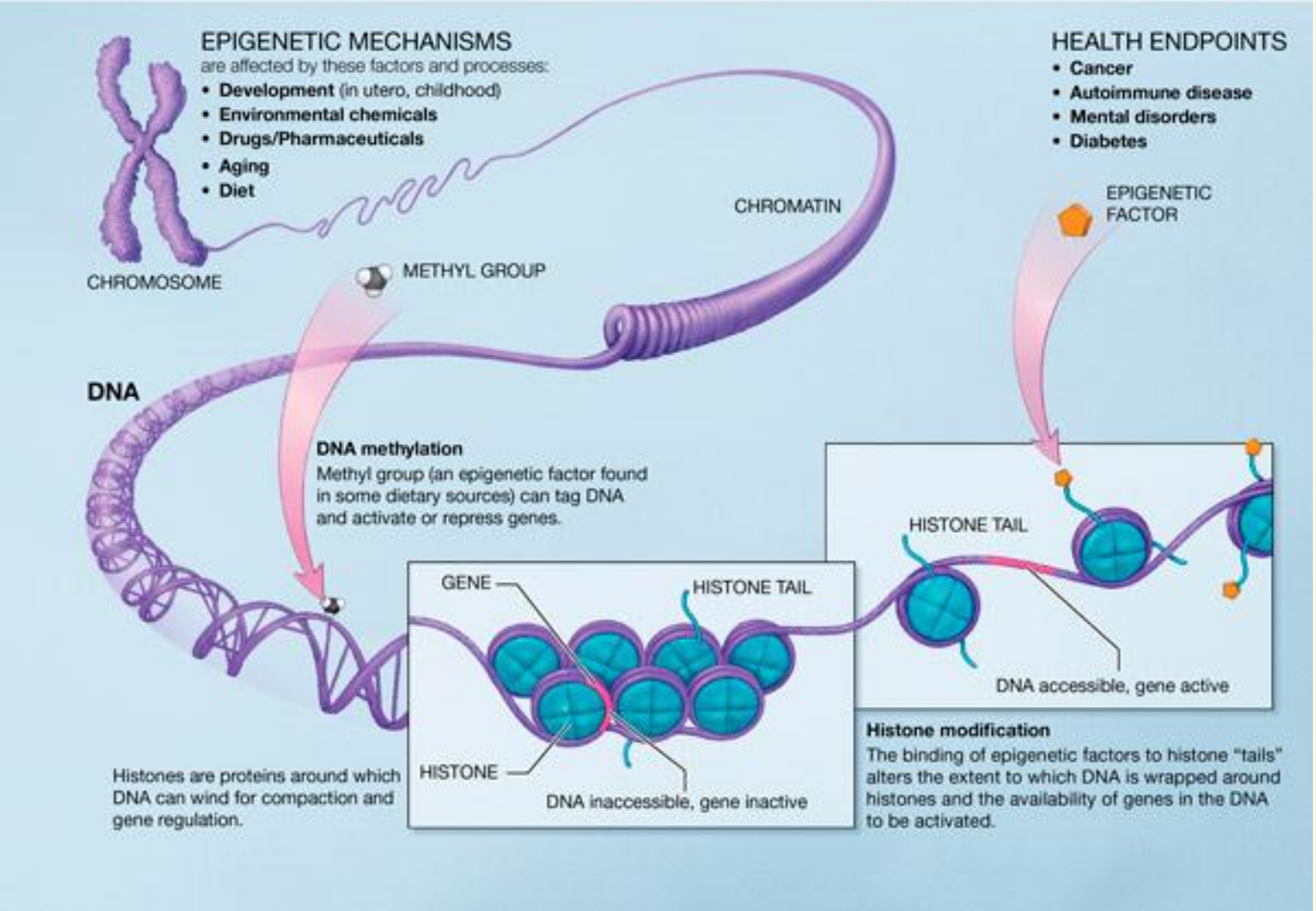
Post-translational modifications of proteins :

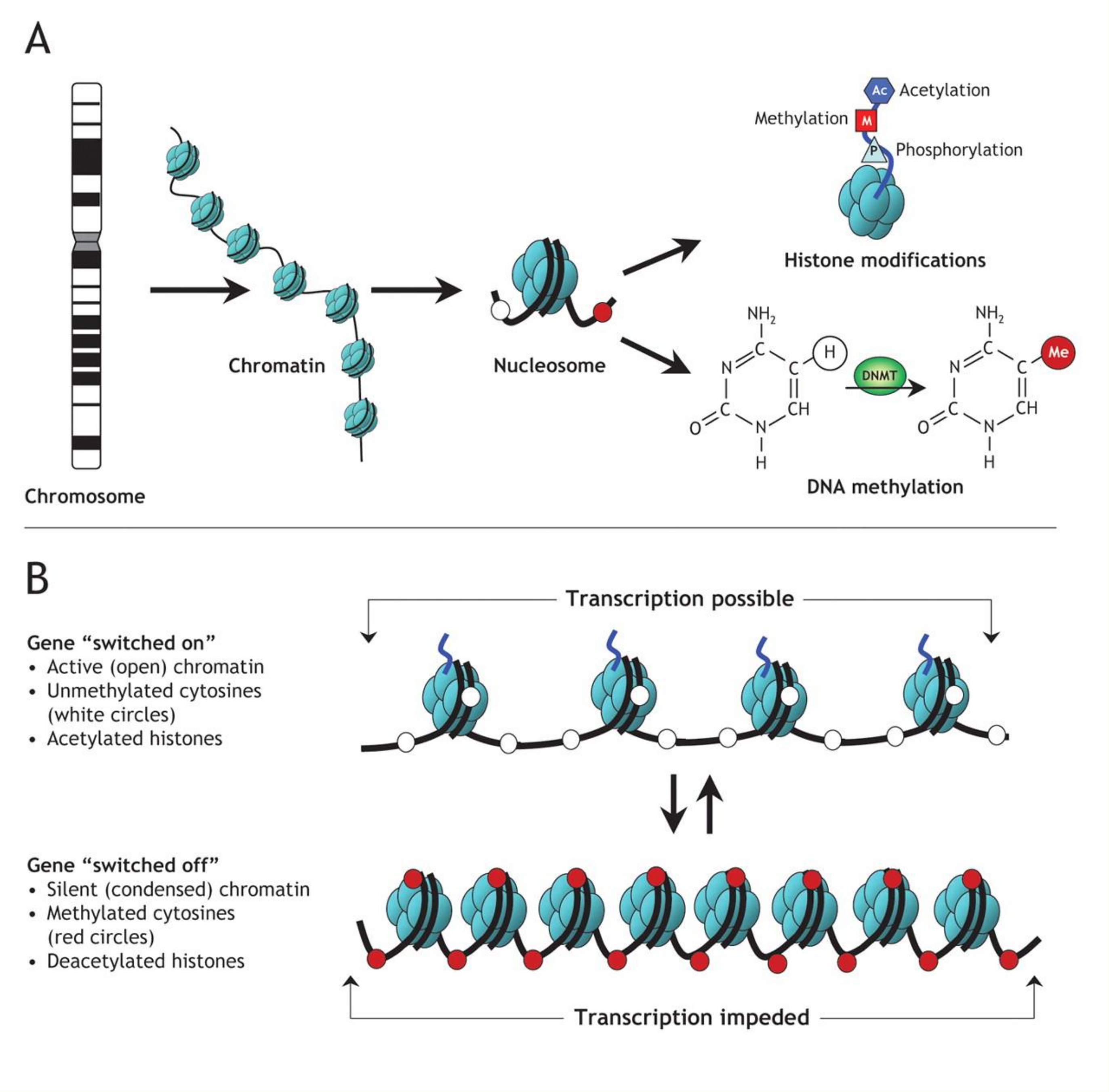
- extends functionality by attaching other groups (e.g. acetate)
- changes chemical nature of amino acids.
- structural changes (disulfide bridges).
- enzymes may remove amino acids or cut the peptide chain in the middle.

s



Epigenetic information transfer

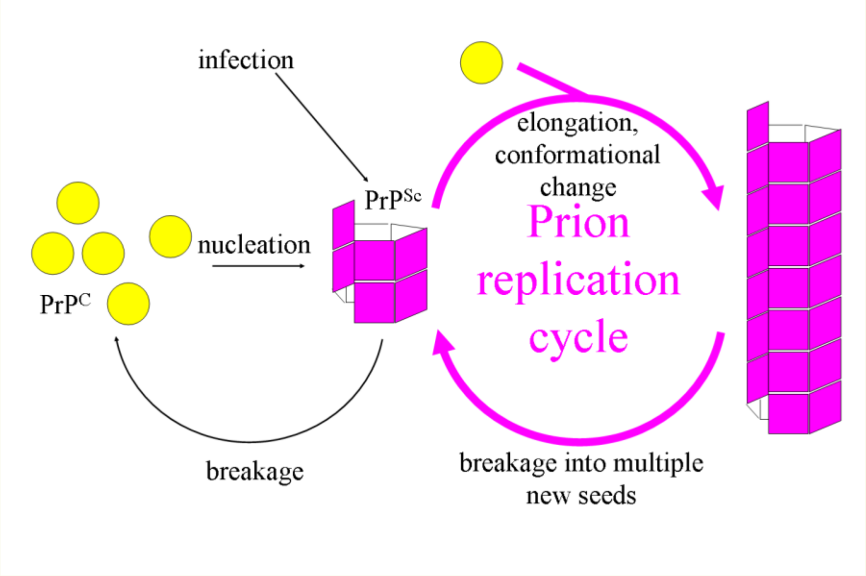




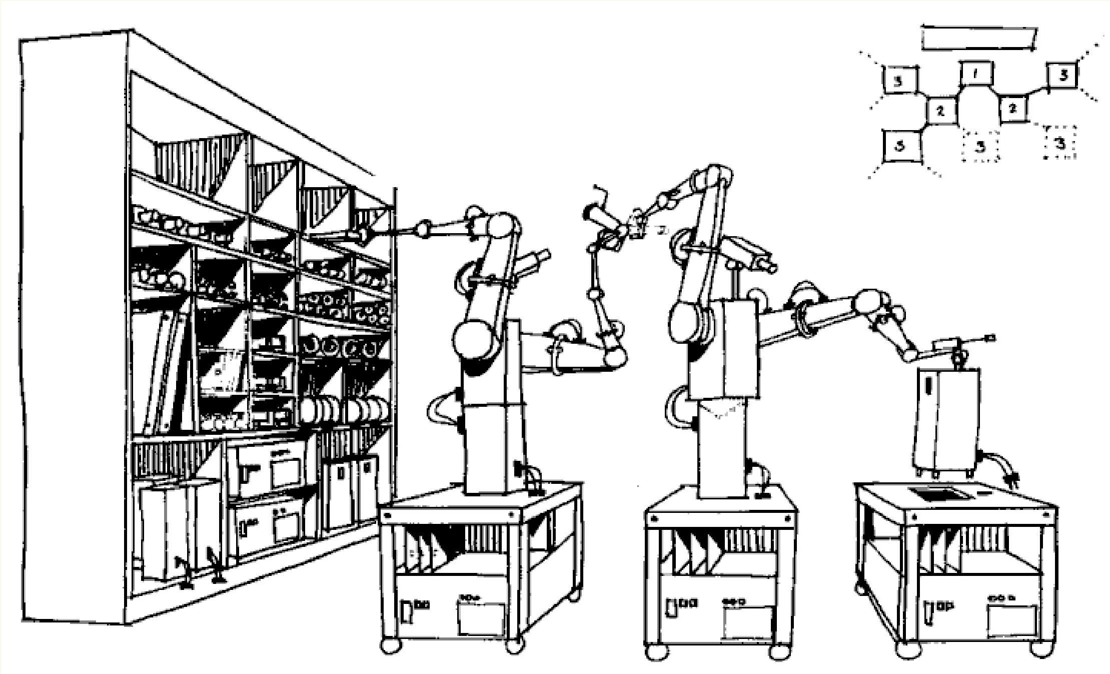
A
Acetylation
Methylation
Phosphorylation
Ac
M
P
Histone modifications
Chromatin
Nucleosome
NH2
H
DNMT
Me
DNA methylation
Chromosome
B
Transcription possible
Gene “switched on”
• Active (open) chromatin
• Unmethylated cytosines (white circles)
• Acetylated histones
Gene “switched off”
• Silent (condensed) chromatin
• Methylated cytosines (red circles)
• Deacetylated histones
Transcription impeded

Prions transfer folding state

- Prions propagate by transmitting mis-folded state.
- **Chain reaction**: conversion of properly folded proteins to prion form.
- **Amyloid fold**: polymer of tightly packed beta sheets
- Amyloids are fibrils grow at their ends and replicating by breaking.



Information transfer by inheritance: Self-replication



Proposed demonstration of simple robot self-replication,

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

Von Neumann's universal constructor

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

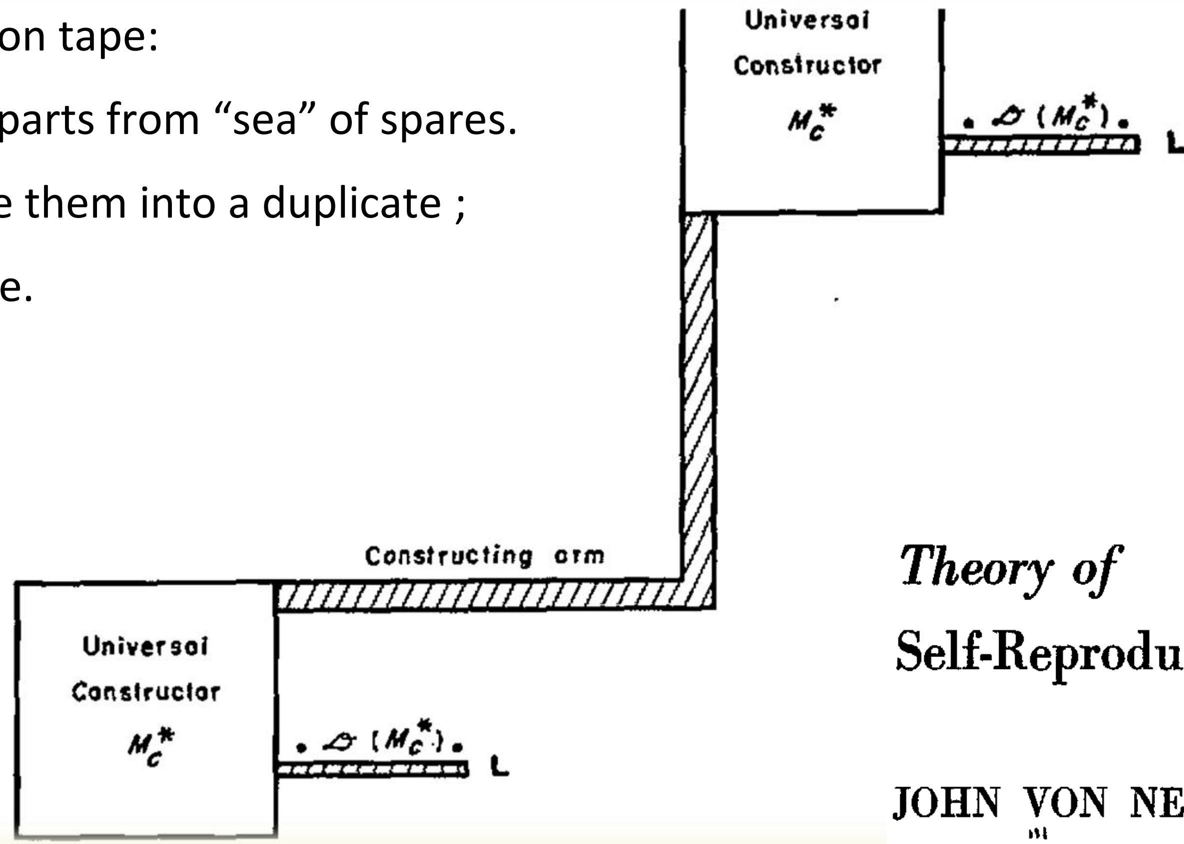
- Program on tape:

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

(ii) assemble them into a duplicate ;

(iii) copy tape.

.



Theory of

Self-Reproducing Automata

(1966)

JOHN VON NEUMANN

edited and completed by Arthur W. Burks

Von Neumann's design allows open-ended evolution

Motivated by biological self-replication:

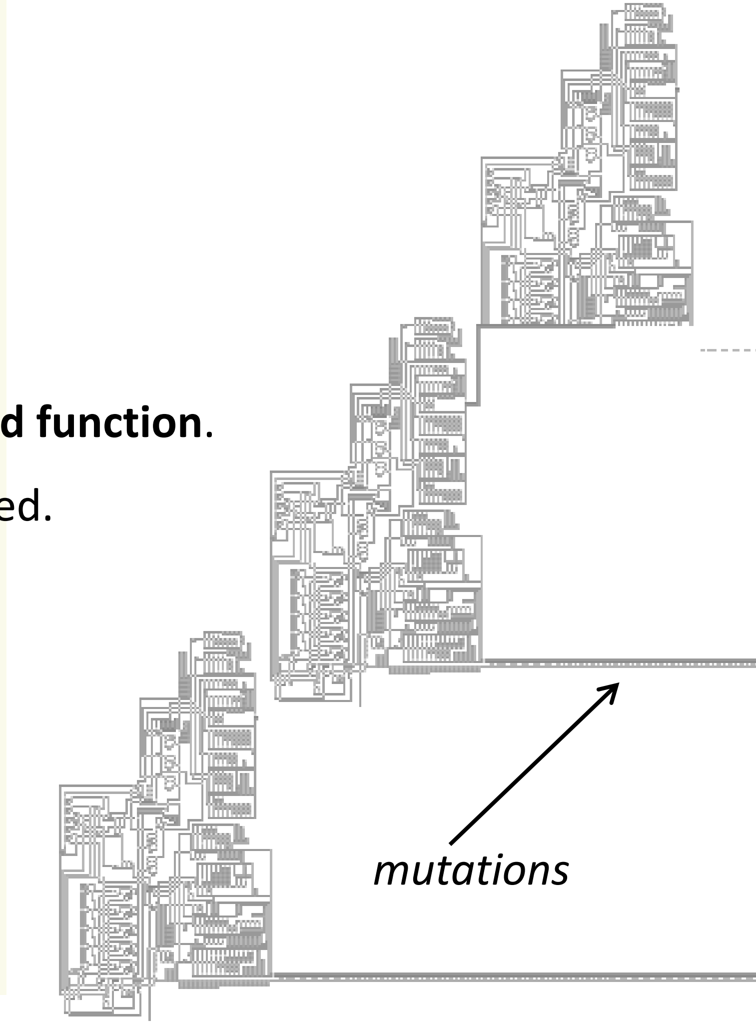
- Construction universality.
- Evolvability.

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

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

How to design

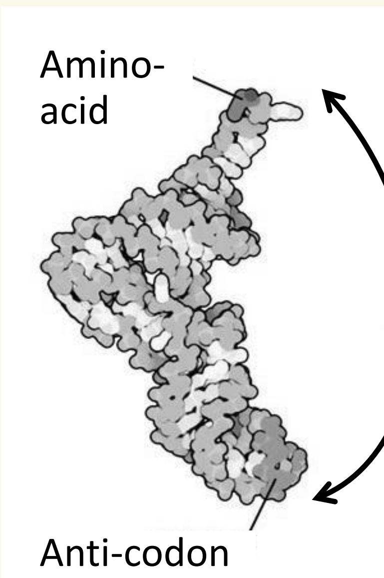
fast/accurate/compact constructor?



Implementation by Nobili & Pesavento (1995)

Ribosomes translate nucleic bases to amino acids

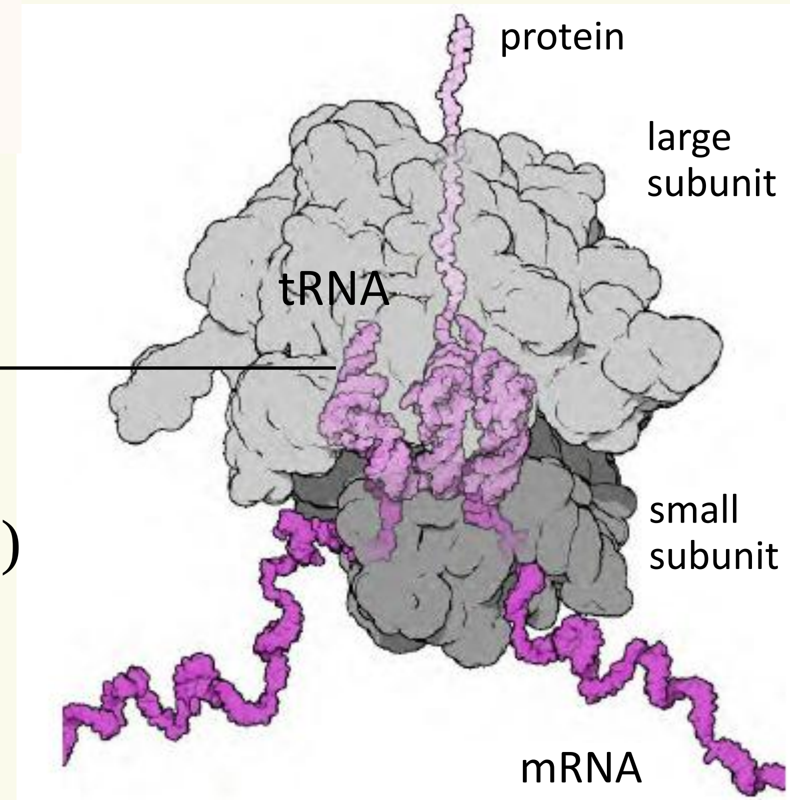
- **Ribosomes** are *large* molecular machines that synthesize proteins with **mRNA** blueprint and **tRNA**s that carry the genetic code.



tRNA

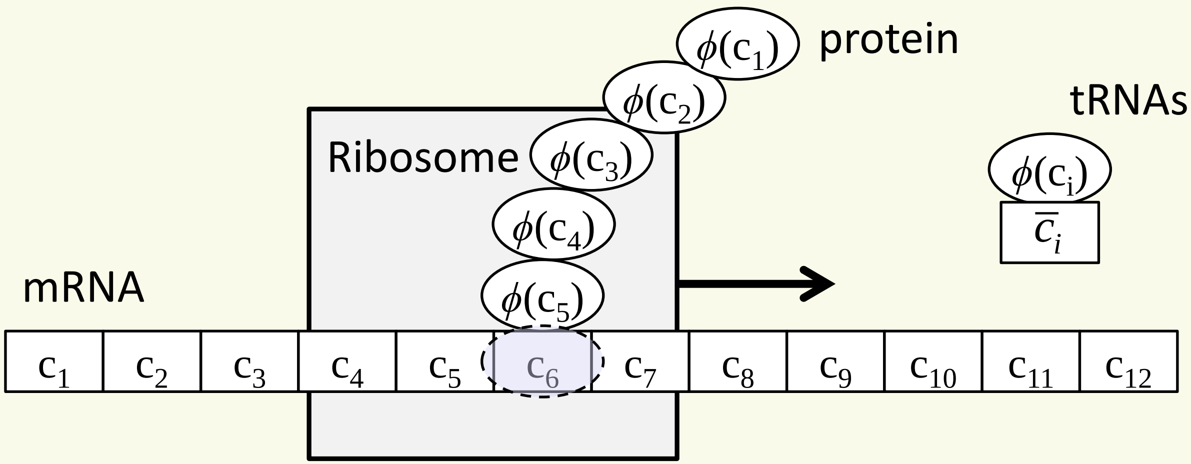
genetic code

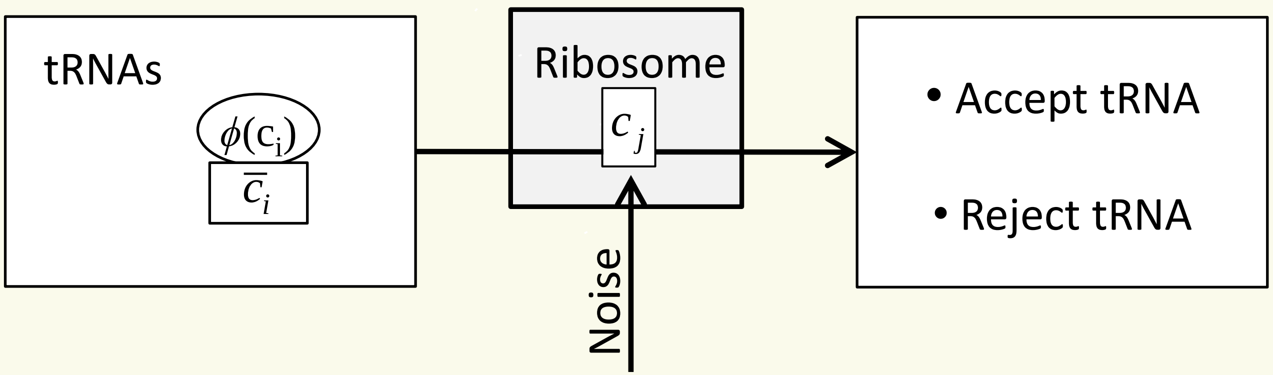
$$\text{amino-acid} = \phi(\text{codon})$$



Goodsell, *The Machinery of Life*

Ribosome needs to recognize the correct tRNA



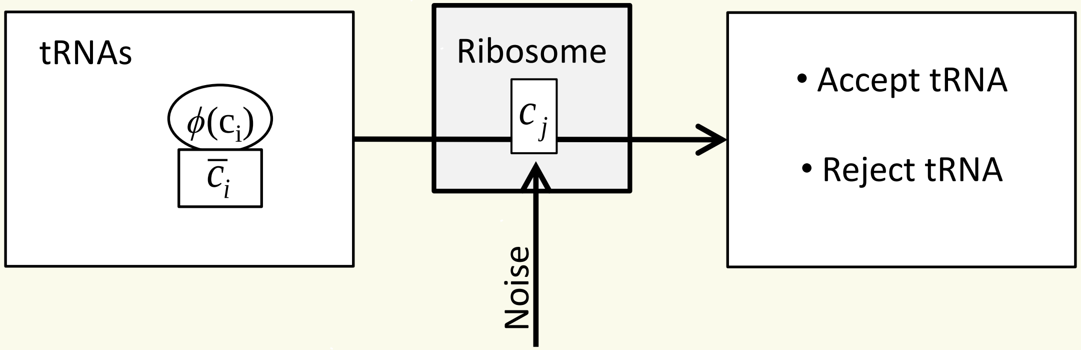


(i) binding wrong tRNAs: amino-acid $\neq \phi(\text{codon})$

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

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

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?

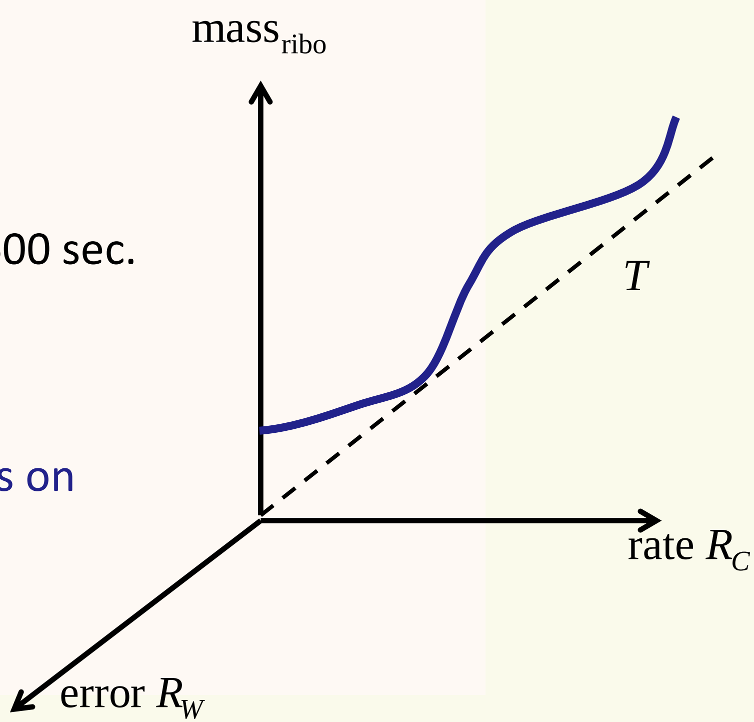
Ribosome sets physical limit on self-reproduction rate

Large fraction of cell mass is ribosomes.

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

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

- Fastest growing bacteria (*Clostridium perfringens*): *T* ~ 600 sec.



Problem: how ribosome accuracy affects fitness depends on

(i) Basic protein properties (mutations).

(ii) Biological context (environment etc.).

Challenge of molecular coding

Quality (Distortion)

- Molecular recognition in a noisy, crowded milieu.
- Many competing *lookalikes*.
- Weak recognition interactions ~ $k_B T$.

"Synthesis of reliable organisms from unreliable components"
(von Neumann, Automata Stud., 1956)

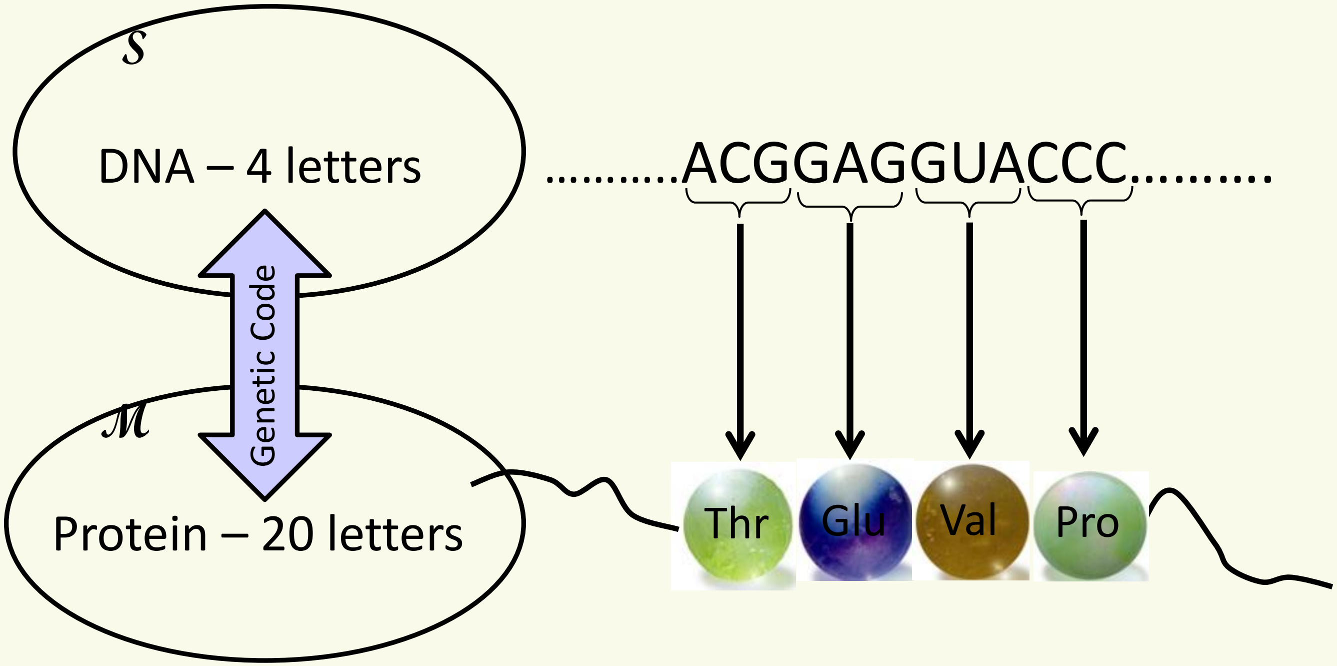
Cost (Rate)

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

Rate-distortion theory (Shannon 1956)

D Goodsell

The genetic code is main information 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.

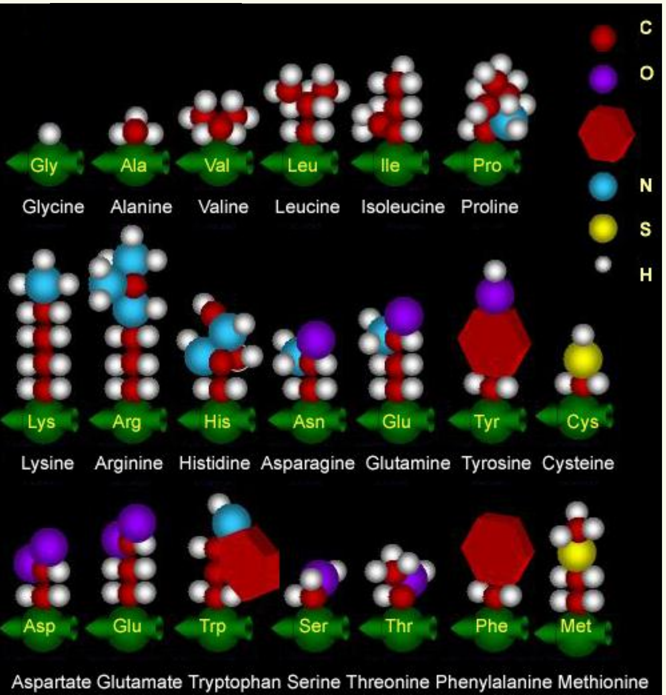
- Proteins are amino acid polymers which perform most biological functions

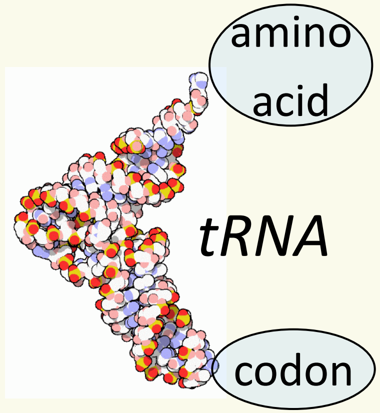
- Diversity of amino-acids is essential to protein functionality.

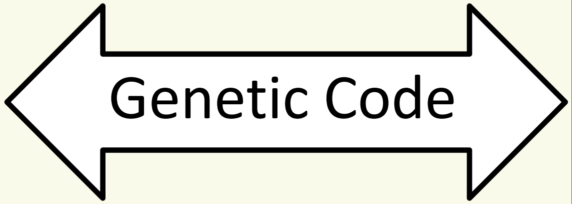
The genetic code maps codons to amino-acids

- 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







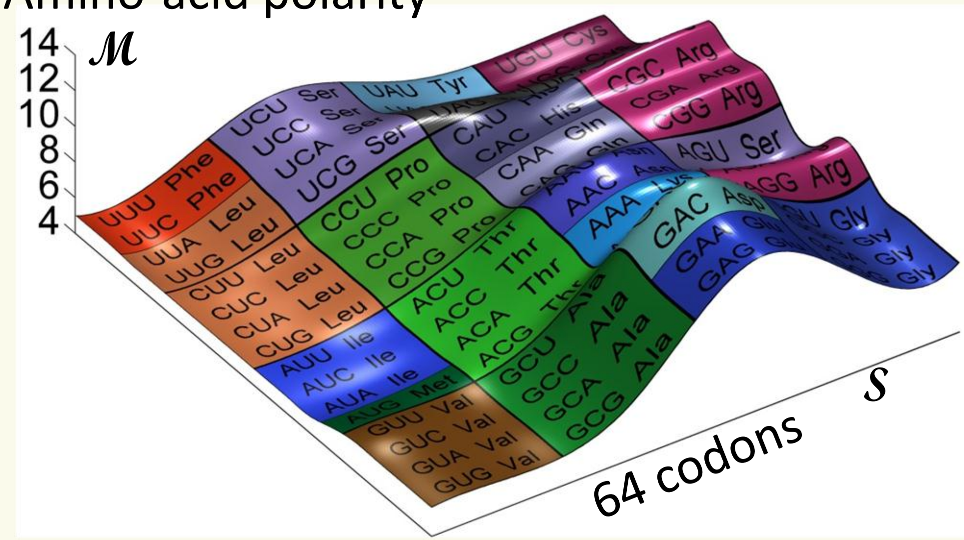
64 codons

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

Amino-acid polarity



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

Generic properties of molecular codes?

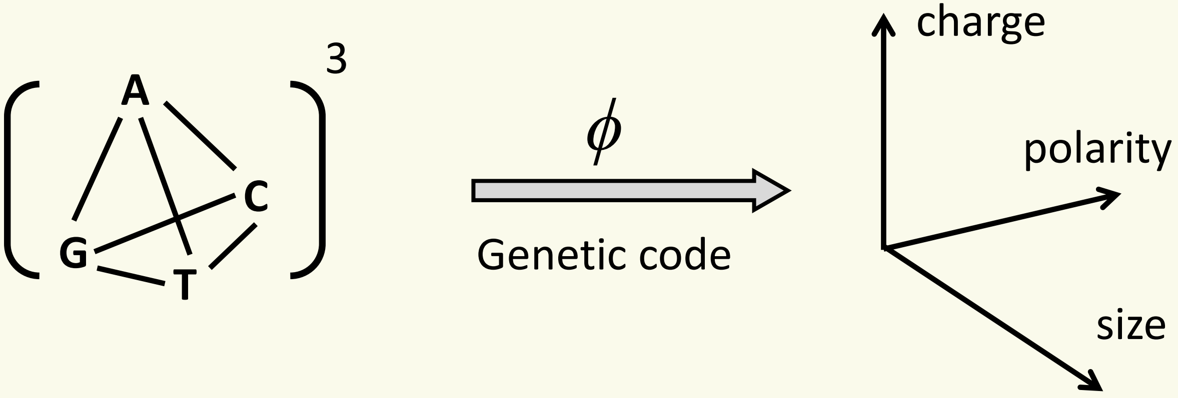
-- to withstand noise at a minimal cost.

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.

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

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

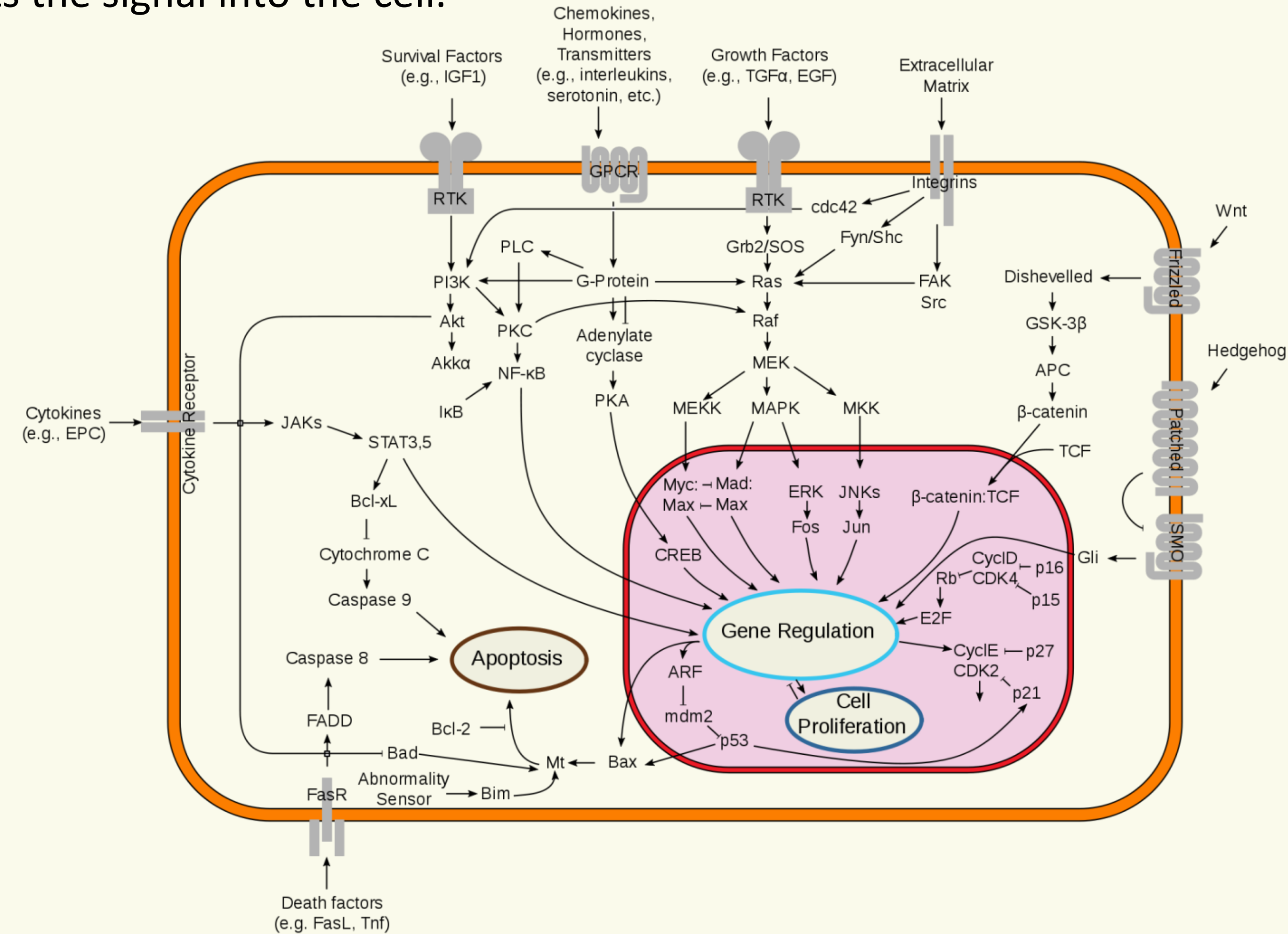


- **Translation machinery**, whose main component is the **ribosome**, facilitates the map.

Signal transduction

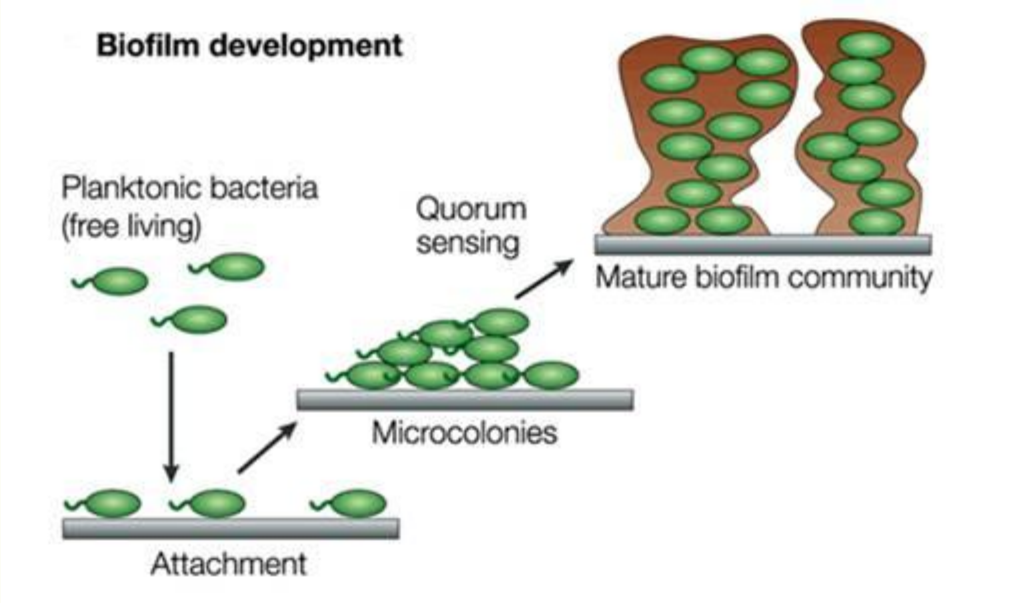
Signal transduction:

- signaling molecule activates specific receptor on cell membrane.
- messenger transmits the signal into the cell.



Quorum sensing

- **Quorum sensing:** stimulus and response correlated to population density.
- Many bacteria use quorum sensing to vary gene expression according to the density of their local population.
- Some social insects use quorum sensing to locate the nest.
- Used in robot flocks.



Living Information - Overview: molecules, neurons, population and evolution

Living systems as information processors:

- *Sources of information in Life*: sequential information, cell composition, environment, population composition.
- *Living information channels*: (modernized) central dogma, replication, codes, receptors – signaling pathways, population dynamics, quorum sensing.
- *Information processing*: circuits and their elements, neural networks, genetic networks.
- *Information output*: transcription level, decisions, cell fate, differentiation and development; feedback (information loop).