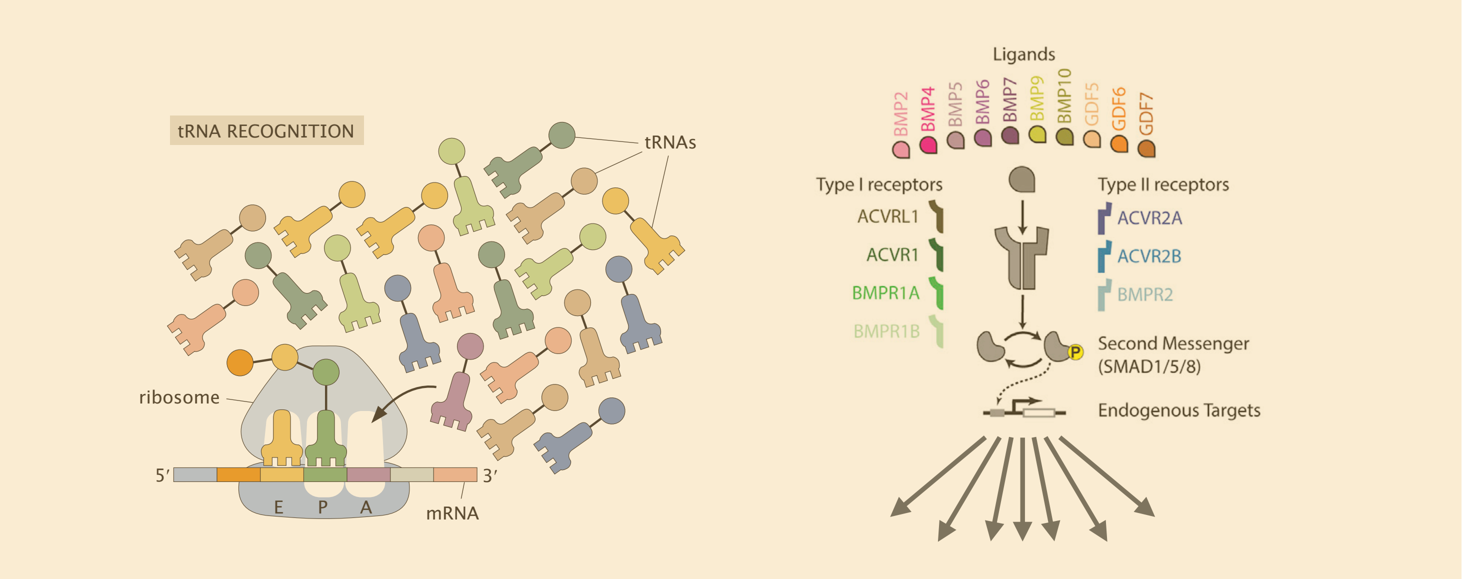
What is biological information?



Course 2: Biological codes

Thomas Lecuit

chaire: Dynamiques du vivant

COLLÈGE DE FRANCE
1530

Plan

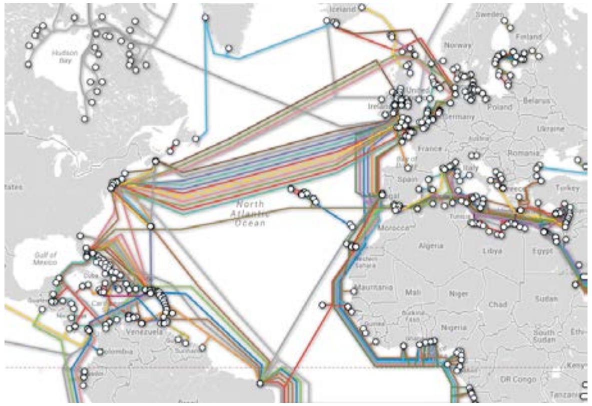
1. General features of chemical information encoding and decoding
2. Case study 1: The genetic code
3. Case study 2: Transcriptional regulatory code
4. Case study 3: Signalling codes
5. Case study 4: Adhesion codes
6. Conclusions

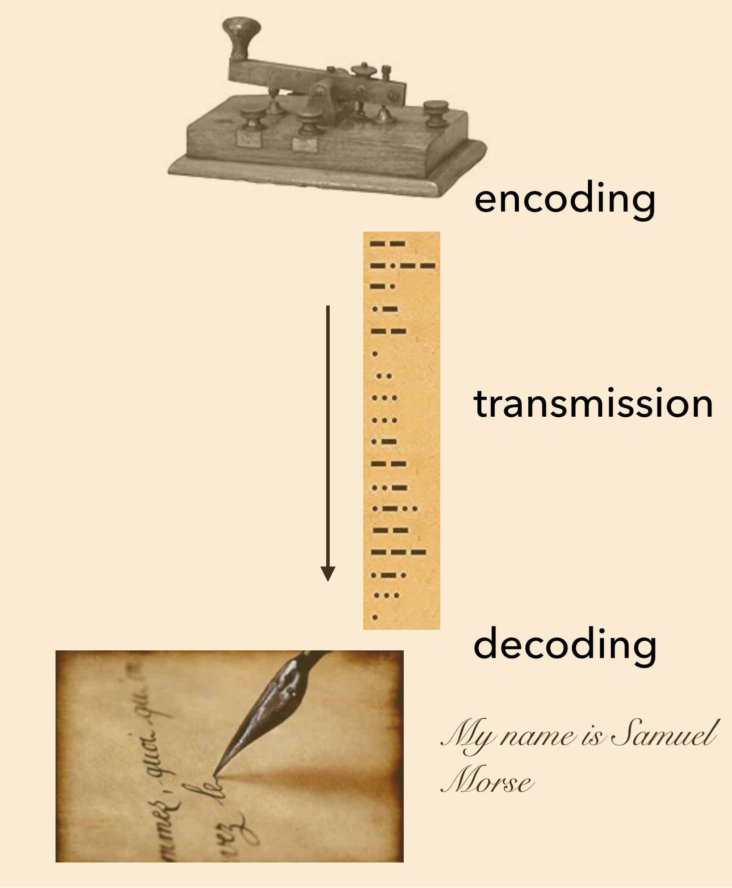
Communication, Information, Codes in Humans

- Information is:

1. Encoded
2. Sent (sender)
3. Transmitted (via electric signals)
4. Interpreted (receiver)

- Information flows:



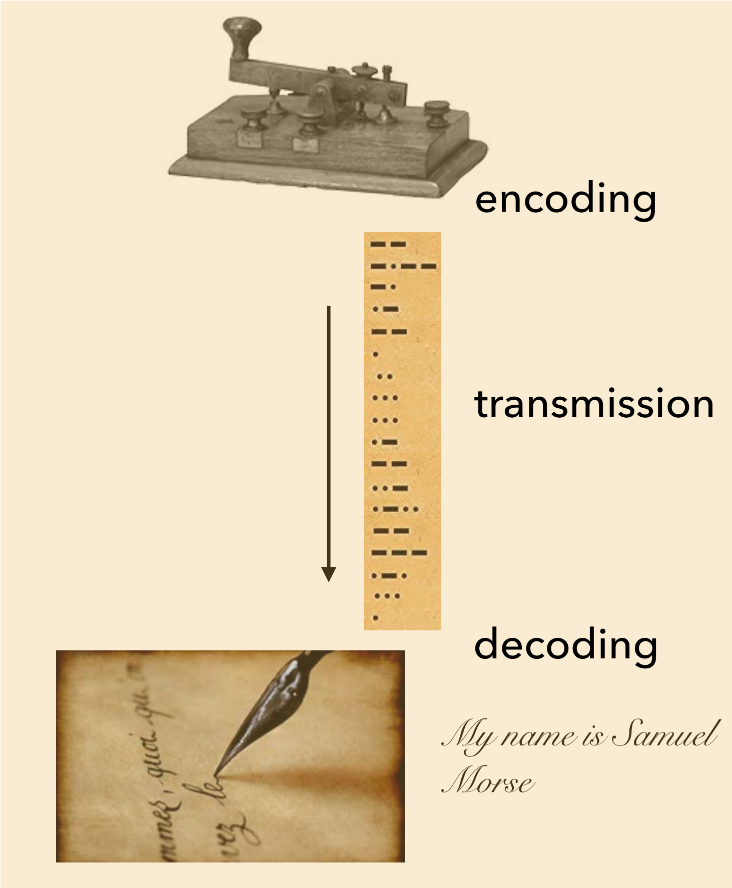


Samuel Morse (1791-1872)

Communication, Information, Codes in Humans

- Information is:
  1. Encoded
  2. Sent (sender)
  3. Transmitted (via electric signals)
  4. Interpreted (receiver)

- A code is used as an *intermediate* between two forms of information
- A code *transforms* an information into another.
- In other words, a code changes a *representation* into another one.



encoding

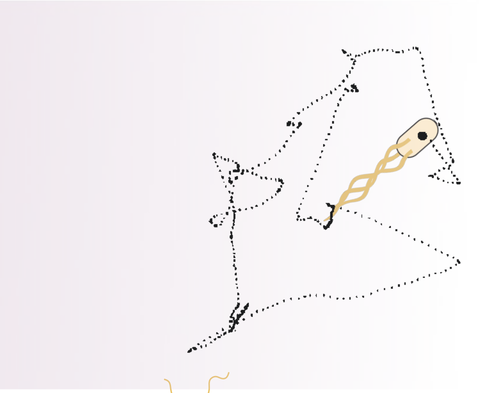
transmission

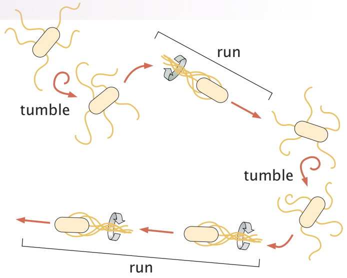
decoding

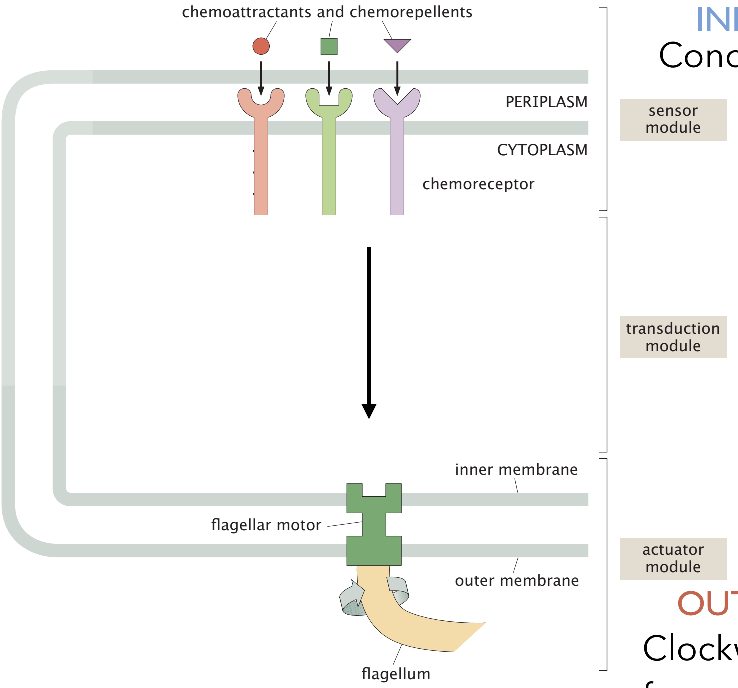
My name is Samuel Morse

Biological Information is mostly chemical

Reading and decoding information from the environment during chemotaxis

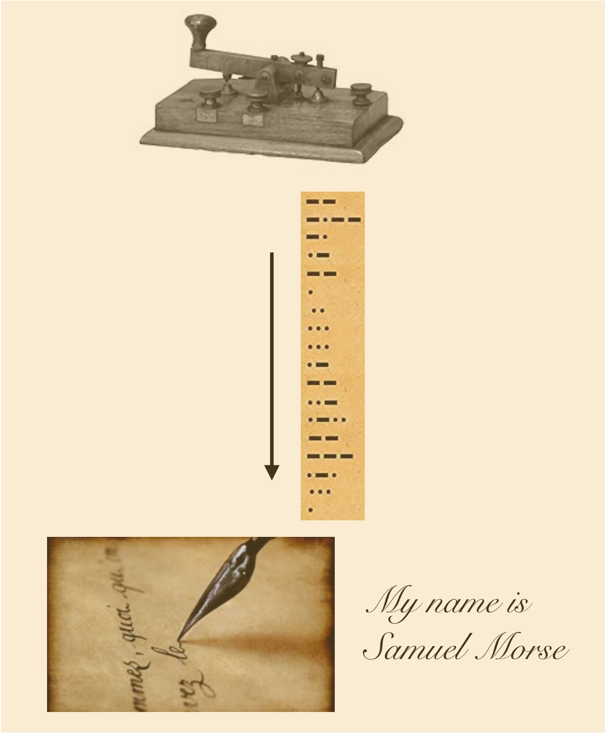






INPUT
Concentration

OUTPUT
Clockwise rotation frequency

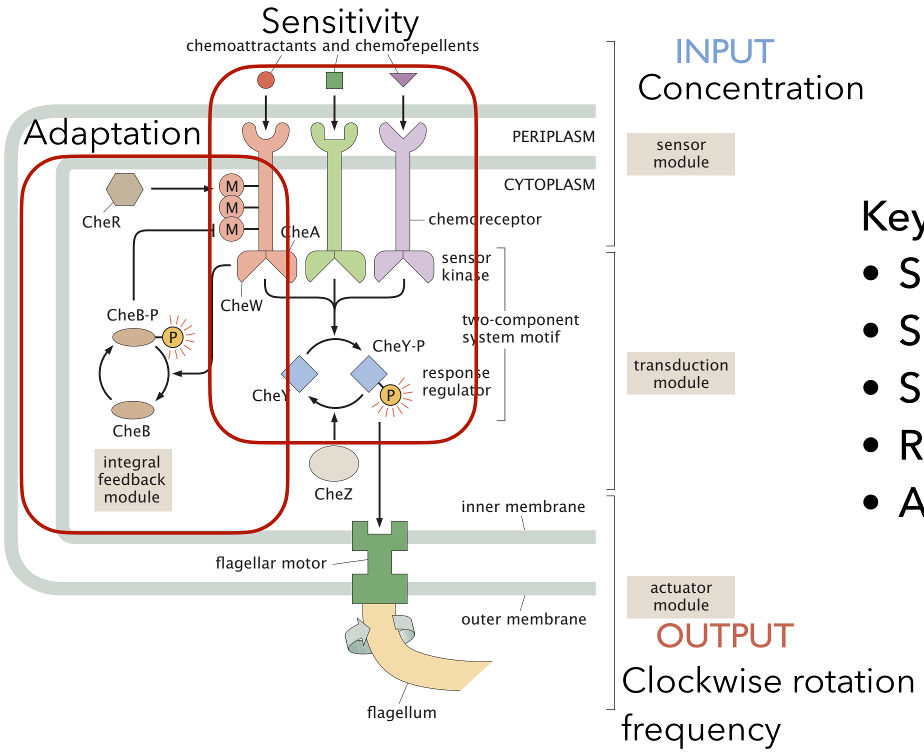


- Moving up the gradient reduces the tumbling frequency
- Cells spend more time going up the gradient than down, so they go up the gradient

R. Phillips, The Molecular Switch: signaling and allostery. *Princeton Univ. Press*. 2020

Information is mostly chemical

Reading and decoding information from the environment during chemotaxis



Key features:

- Specificity
- Sensitivity/Gain
- Speed
- Resistance to noise
- Adaptation

- **Specificity:** glucose, ribose, galactose, aspartate, serine

- **Sensitivity**

  A ramp that increases the receptor occupancy by as little as 1 molecule/second (1 part in 600 Tar receptors/cell, or 0.0016) leads to a steady state increase in flagellar rotational bias by ~0.1).

  This corresponds to a change in run length by a factor of ~3.

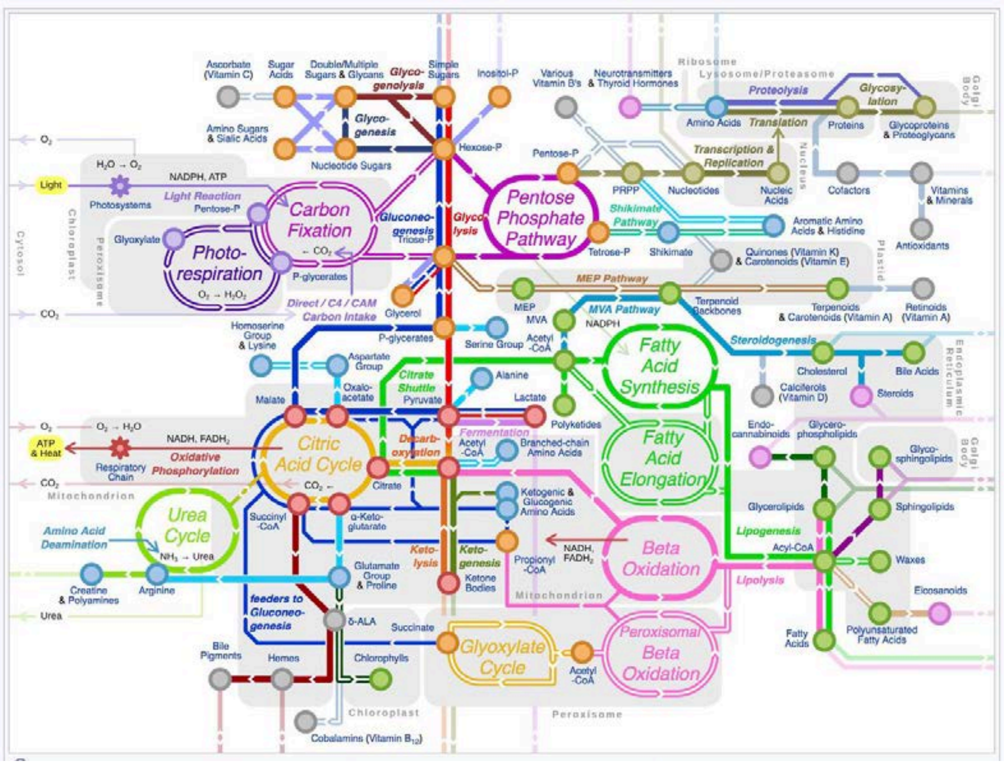
J. Segall, SM Block and & HC. Berg, *PNAS* 83, 8987-8991 (1986).

- **Adaptation:** resetting in a gradient and large amplitude

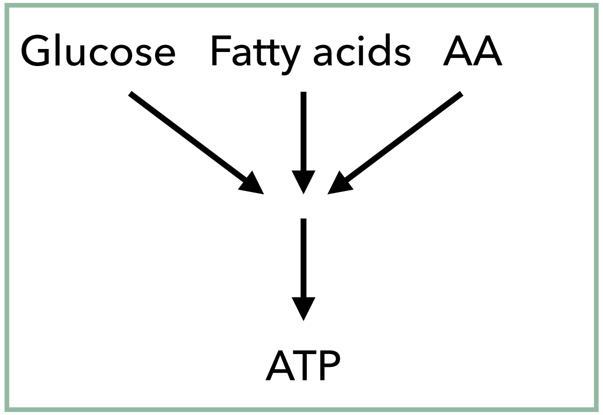
Information is mostly chemical

And based on molecular recognition: Metabolic pathways

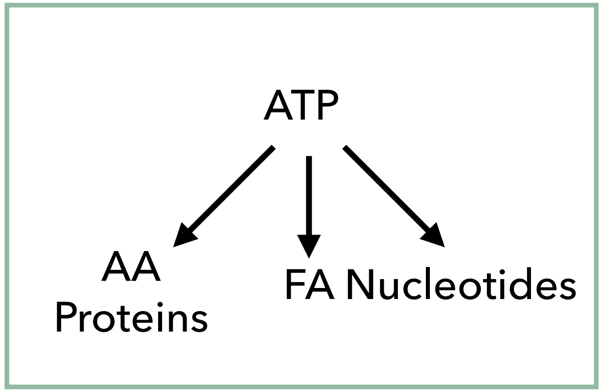
Enzyme/substrate, Regulator/Enzyme



https://en.wikipedia.org/wiki/Template:Metabolic_metro


Glucose Fatty acids AA

ATP

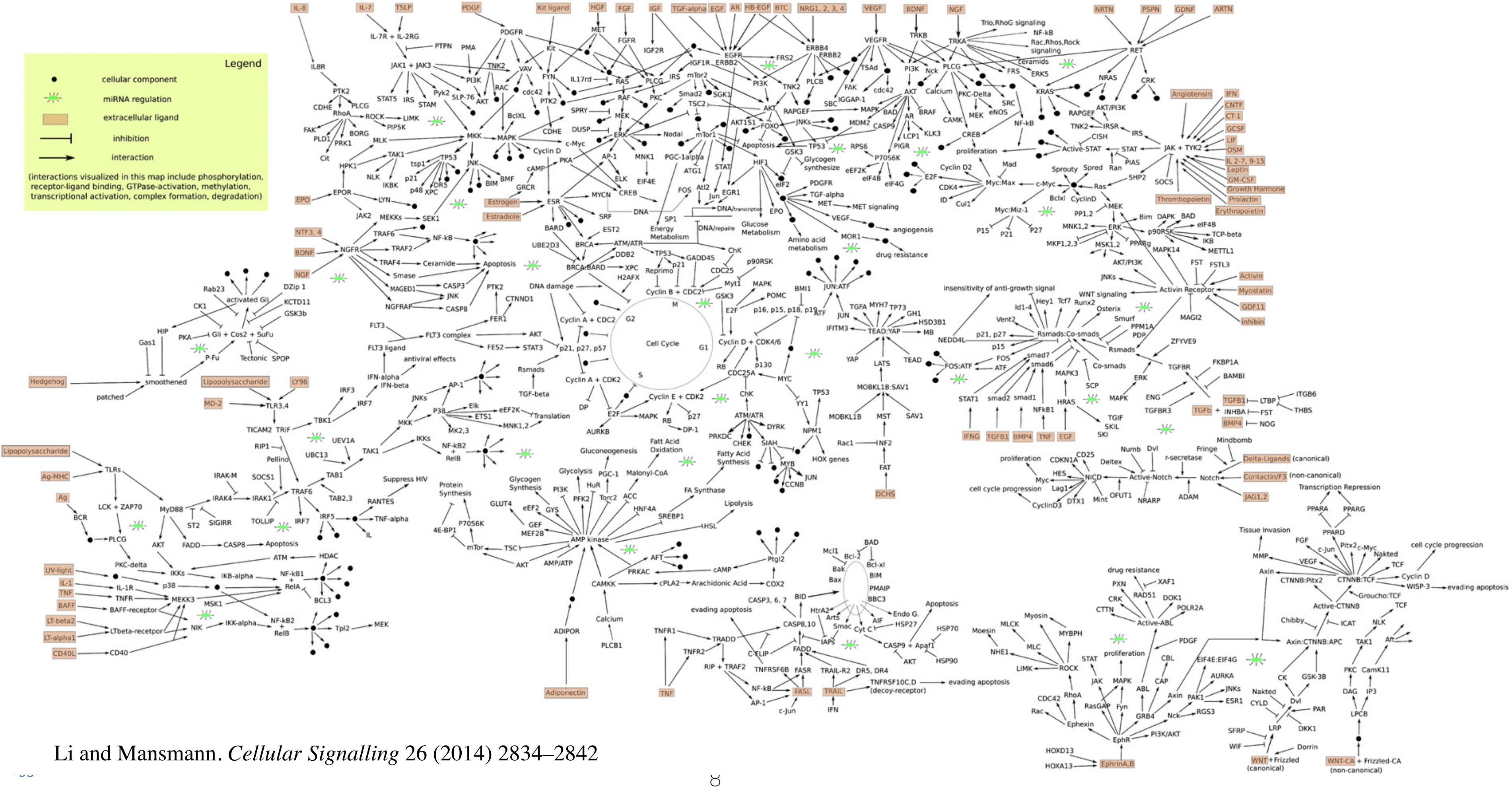

ATP

AA
Proteins

FA Nucleotides

Information is mostly chemical

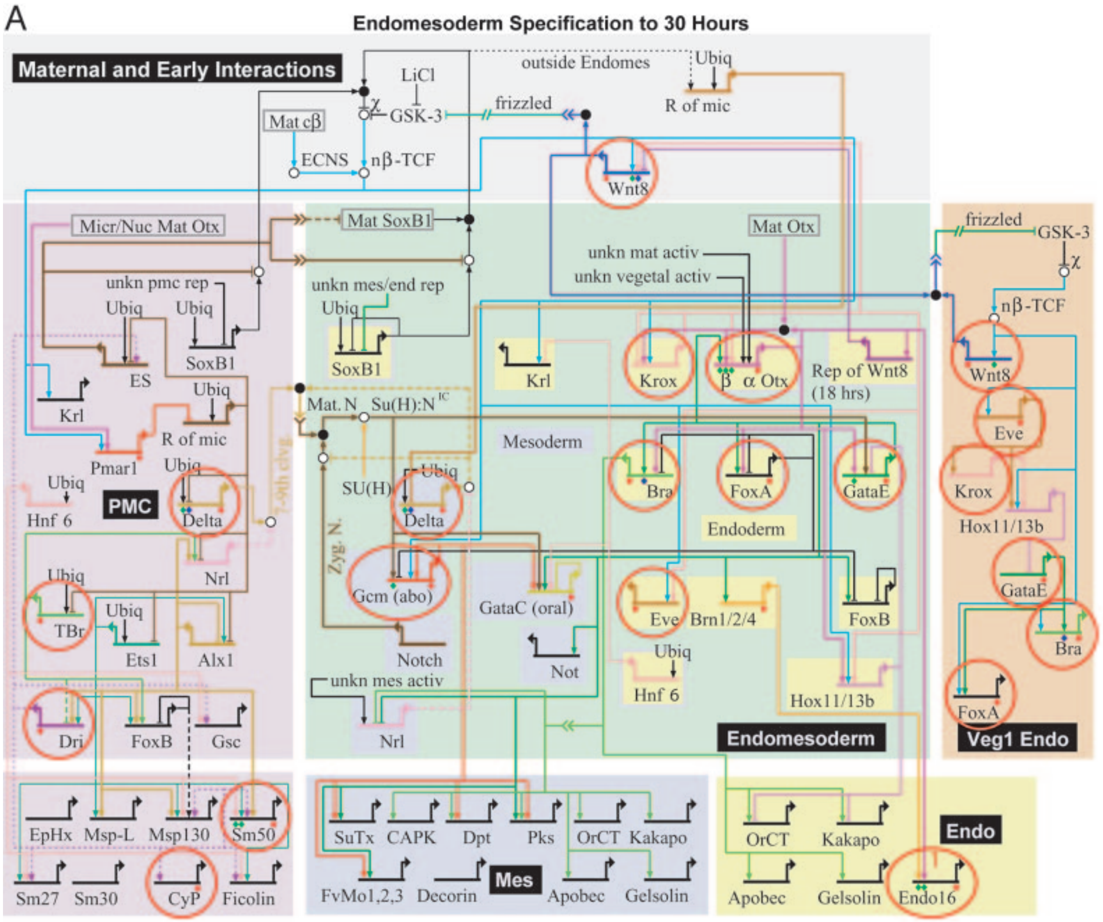
And based on molecular recognition: Signalling pathways



Li and Mansmann. *Cellular Signalling* 26 (2014) 2834–2842

Information is mostly chemical

And based on molecular recognition: Developmental Gene regulatory networks



L. Bodenstein. *Mechanisms of Development*, 162 (2020)
https://doi.org/10.1016/j.mod.2020.103606

But what is information *about* in living systems?

1. **The system and the observer/scientist:**
   - Information underlying the functioning of a system
   - Information used to understand/model/represent a system

2. **Operational definition of information:**
   - Information of a system is the set of parameters and prescriptions that allow an accurate prediction of the system's evolution, given a model.

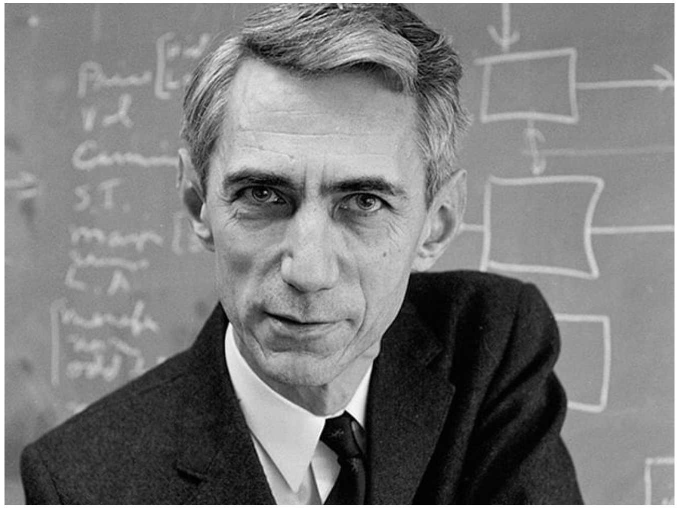
3. **What is relevant or useful information:** (completeness vs sufficiency)

5. **Can information be quantified?**
   - Yes (Shannon, courses #3 and 4) and Not yet (see courses #5 and 6)

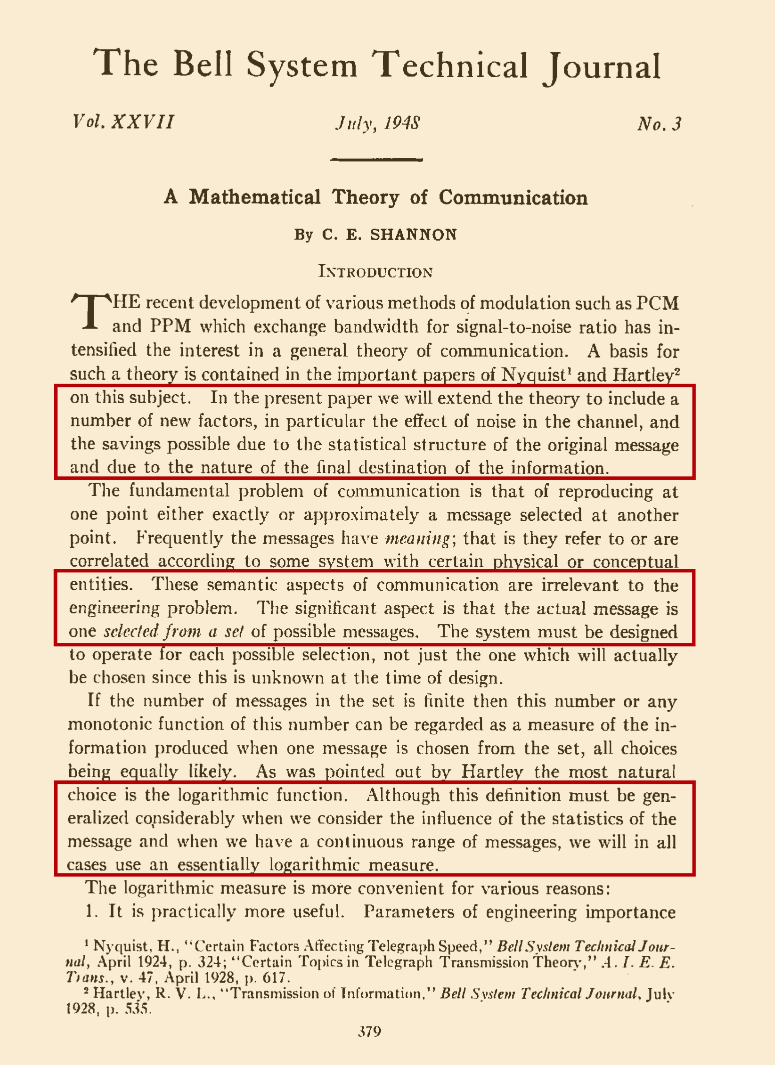
6. **Encoding and decoding information:**
   - simplified (low dimensional) representation of relevant information

Mathematical theory of Information and Communication

- Claude Shannon – 1948
- Key features of information theory:
  - – semantic is not relevant
  - – probabilistic nature of information
  - – considers non uniform frequency of « events » and statistics of the message



Claude Shannon (1916-2001)



The Bell System Technical Journal

Vol. XXVII July, 1948 No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

379

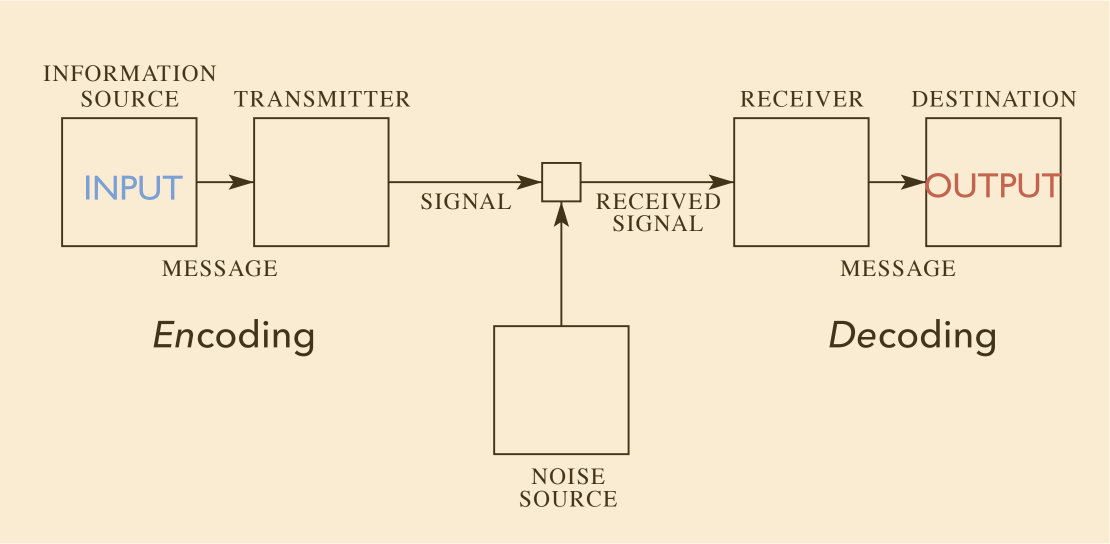
COLLÈGE DE FRANCE 1530

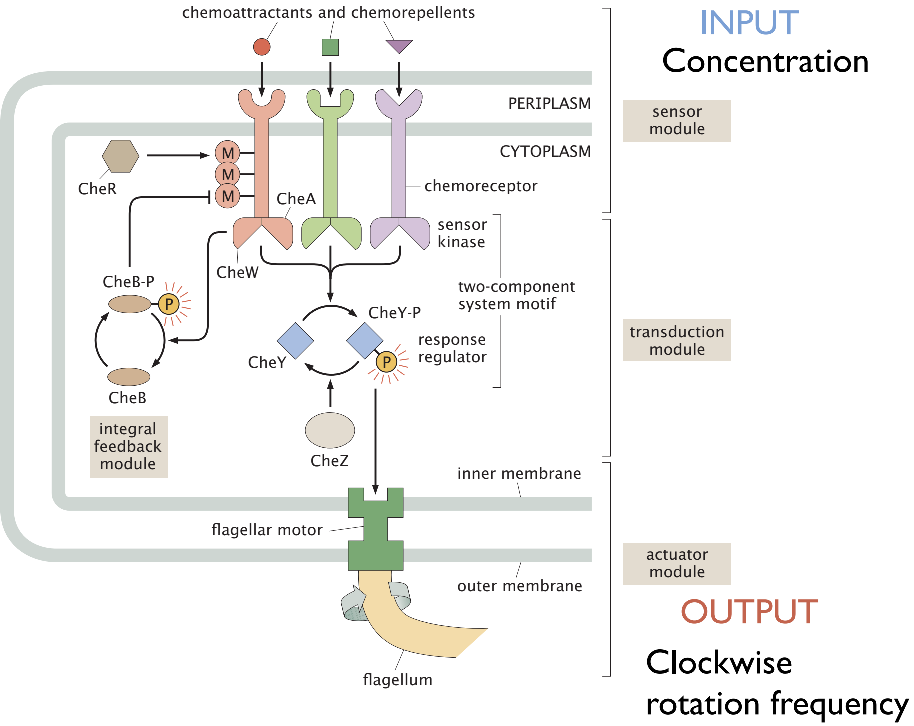
Thomas LECUIT 2024-2025

Theory of Information and Communication

« The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. »

- Basic architecture of *any* communication system
- Information *transfer in a noisy channel*



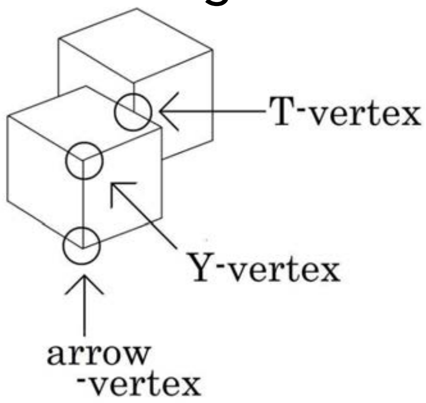


R. Phillips, The Molecular Switch: signaling and allostery. *Princeton Univ. Press*. 2020

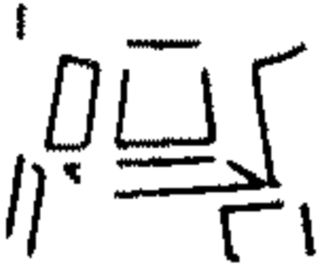
What is relevant information?

Not all pixels in an image have the same relevance, ie. meaningfulness or usefulness

clues for recognition

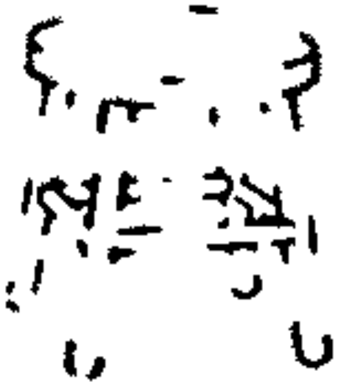


?



Vs

Stool



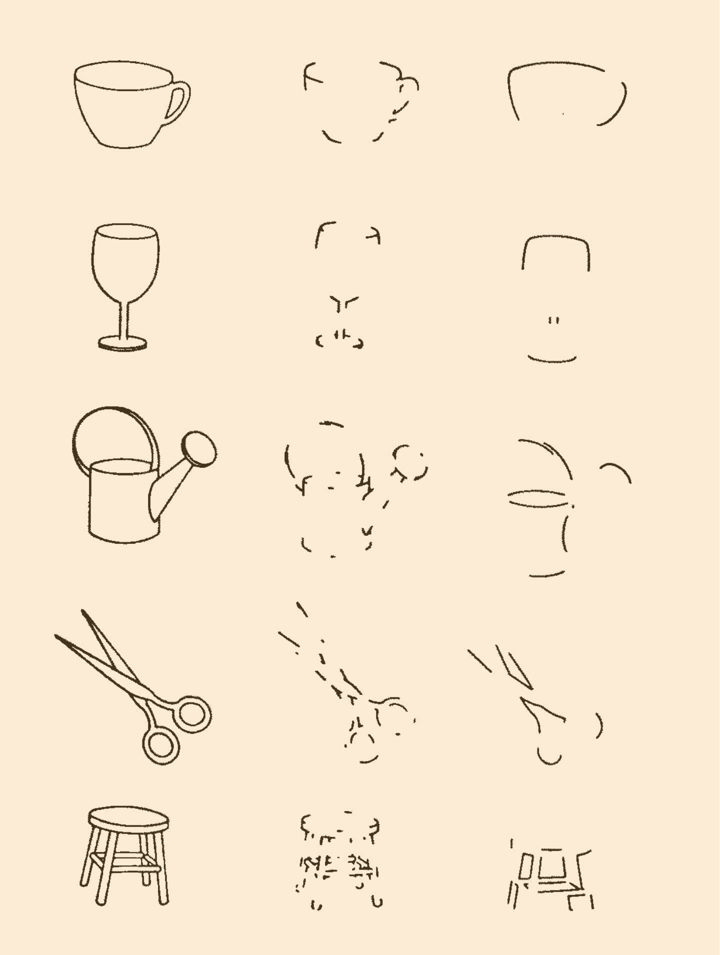
Recognition by components theory

I. Biederman, *Psychological Review* 94, 115–147, 1987

What is relevant information?

THS TXT S NT VR
DFFCLT T RD

Y CN RD THS
F Y TR RL HRD

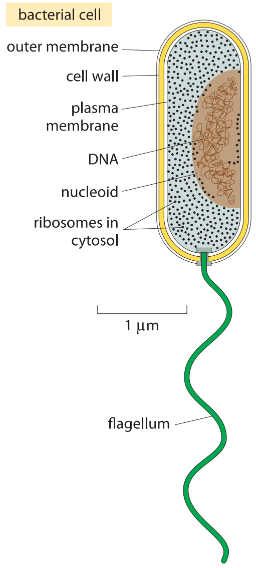


I. Biederman, *Psychological Review* 94, 115–147, 1987

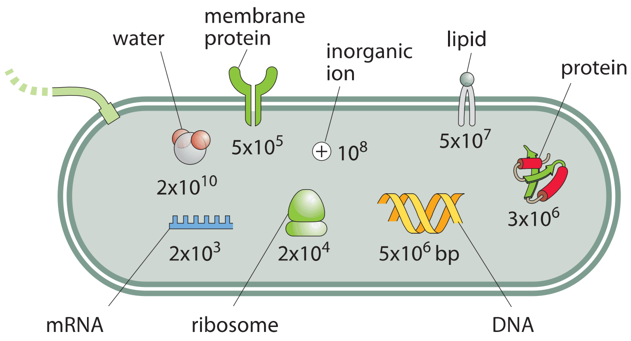
What is relevant biological information?

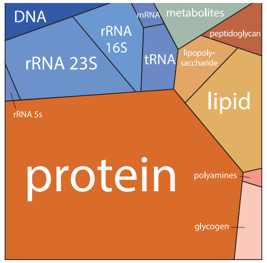
Cells by the numbers

property	*E. coli*	budding yeast	mammalian (HeLa line)
cell volume	0.3–3 µm^3	30–100 µm^3	1,000–10,000 µm^3
proteins per µm^3 cell volume	2–4×10^6		
mRNA per cell	10^3-10^4	10^4–10^5	10^5–10^6
proteins per cell	~10^6	~10^8	~10^{10}

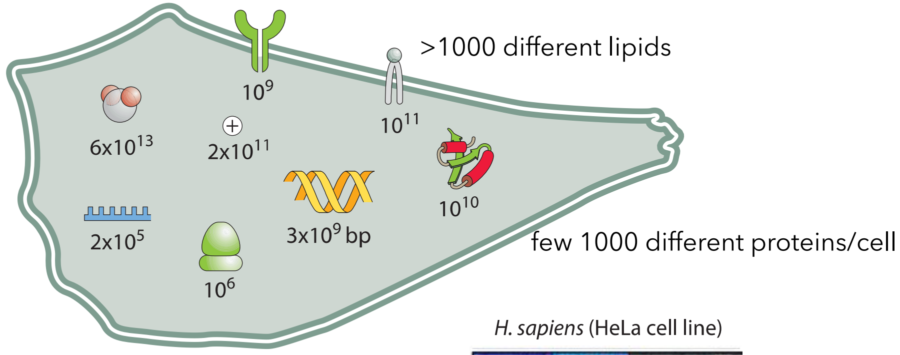


(A) bacterial cell (specifically, *E. coli*: V ≈ 1 µm^3; L ≈ 1 µm; τ ≈ 1 hour)

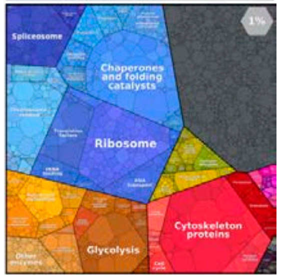




(C) mammalian cell (specifically, HeLa: V ≈ 3000 µm^3; L ≈ 20 µm; τ ≈ 1 day)



H. sapiens (HeLa cell line)



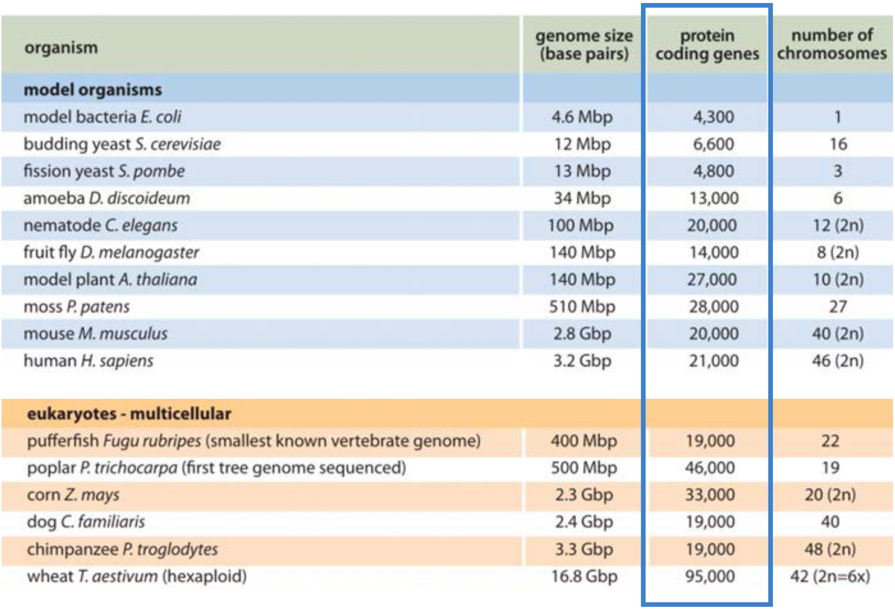
R. Milo and R. Phillips Cell Biology by the numbers. *Garland Science*

Liebmeister et al, R. Milo. *PNAS* (2013) doi/ 10.1073/pnas.1314810111

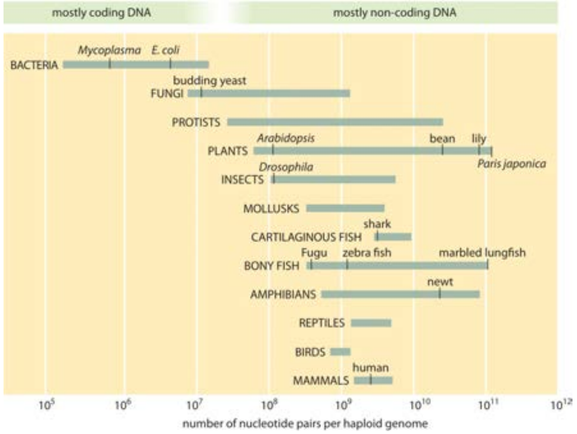
What is relevant biological information?

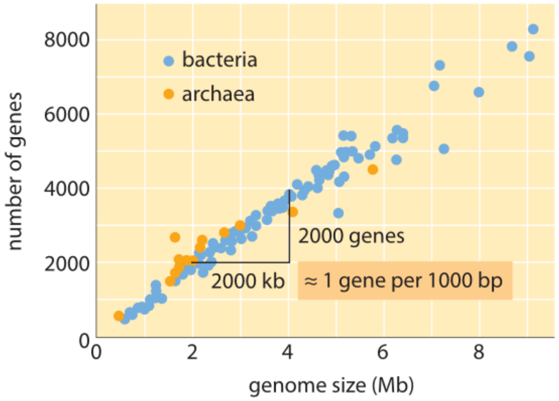
Genome by the numbers

Complexity does not scale with genome size/gene number



organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
eukaryotes - multicellular			
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000	19
corn *Z. mays*	2.3 Gbp	33,000	20 (2n)
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)



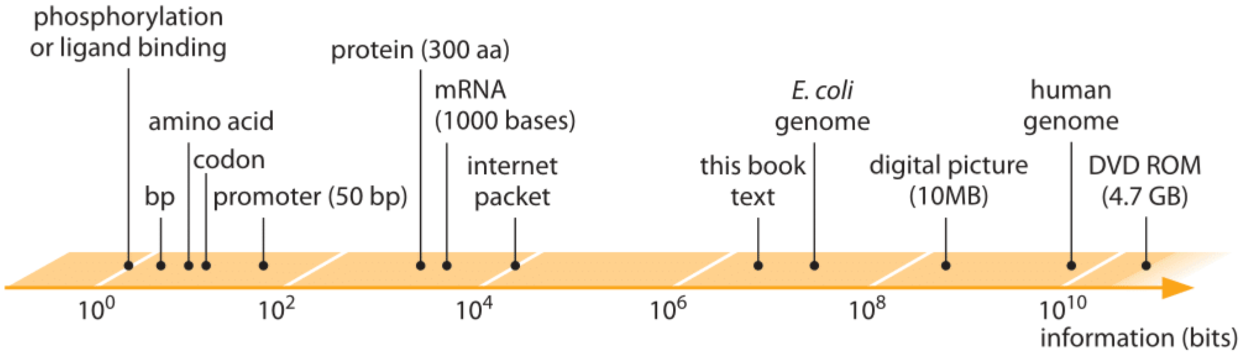


R. Phillips and R. Milo. *Cell biology by the numbers*

What is relevant biological information?

Genome by *bits*

Genomic and chemical information is very large and dense



information, $I = \log_2$ (# possible configurations)

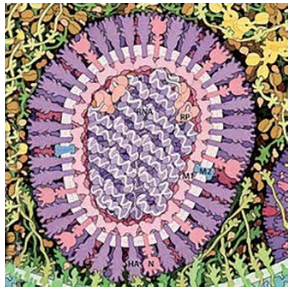
e.g. base pair (bp) has four possibilities ⇒ $I = \log_2(4) = 2$

R. Phillips and R. Milo. *Cell biology by the numbers*

INFORMATION DENSITY OF HARD DRIVE

$$\frac{5 \times 10^{12}\text{ letters}}{100\text{ cm}^3} \approx 5 \times 10^{10}\ \frac{\text{letters}}{\text{cm}^3}$$

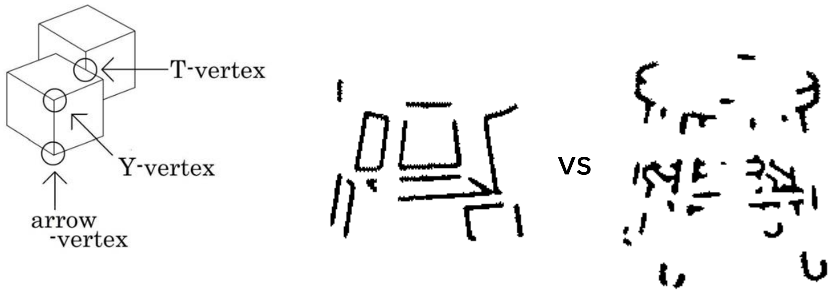
INFORMATION DENSITY OF VIRUS



$$\frac{10{,}000\text{ letters}}{10^5\text{ nm}^3} \approx 10^{20}\ \frac{\text{letters}}{\text{cm}^3}$$

What is relevant biological information?

- Are all biological data/information meaningful to the system itself and to an observer to understand and predict its behaviour to specific endpoints?



- What are the relevant effective tuning variables?
  - Ex: cell actin cortex tension and material properties depend on 100s of proteins
    – few mechanical parameters such as stiffness and viscosity.
    – few molecules with key regulatory functions: MyosinII activation.

Information is mostly chemical

and based on protein affinity/molecular recognition

How to generate/produce a lot from little?

- Key feature: Balancing diversity and specificity

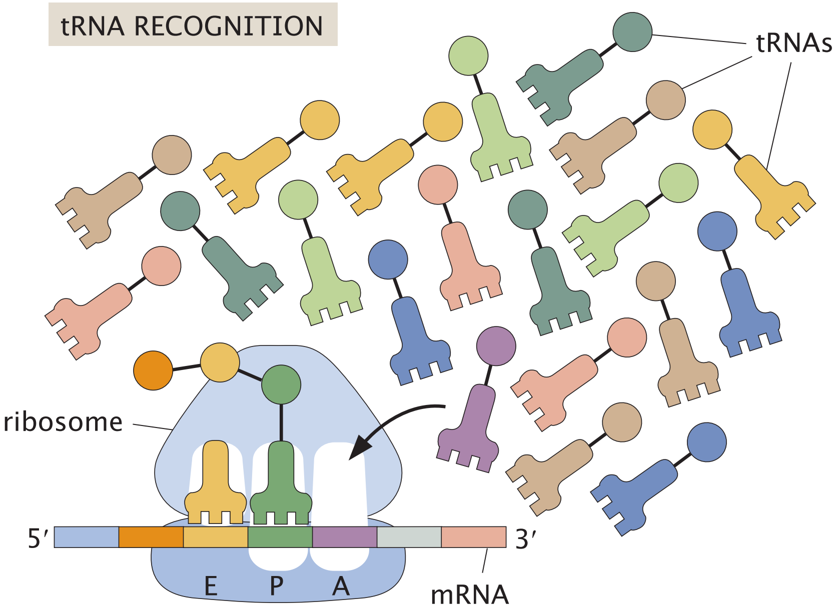
  Increasing diversity can impose a limit on coding system to ensure specificity
- Role of combinatorial properties to increase diversity
- Deterministic encoding vs Encoding in noising/stochastic dynamical systems

- 4 Case studies:
- **tRNA/mRNA (genetic code):** how to encode amino acid recruitment in protein synthesis?
- **TF/DNA (regulatory code):** how to encode gene expression of 10^4 genes &cell state?
- **Ligand/Receptor (signalling code):** how to encode specific signalling responses?
- **CAM/CAM (adhesion code):** how to encode self-organisation of shapes from few 100 CAMs?

Case study 1: The Genetic Code

mRNA/tRNA-aminoacyl:

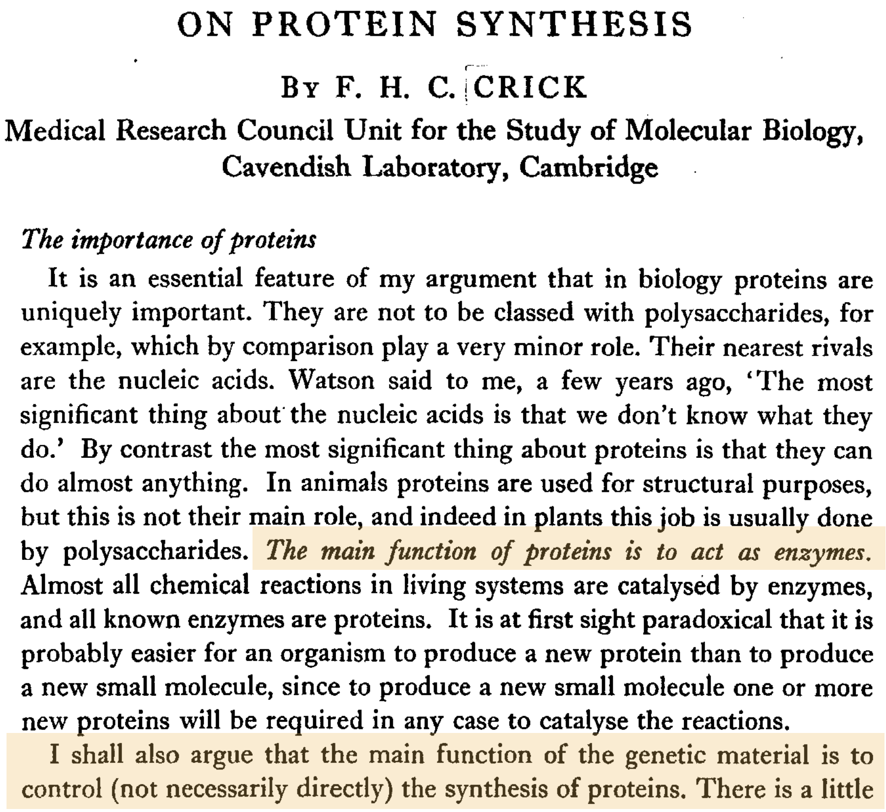
how to encode amino acid recruitment in protein synthesis?



Word code: codon

Message: amino acid

Genetic information flow



ON PROTEIN SYNTHESIS

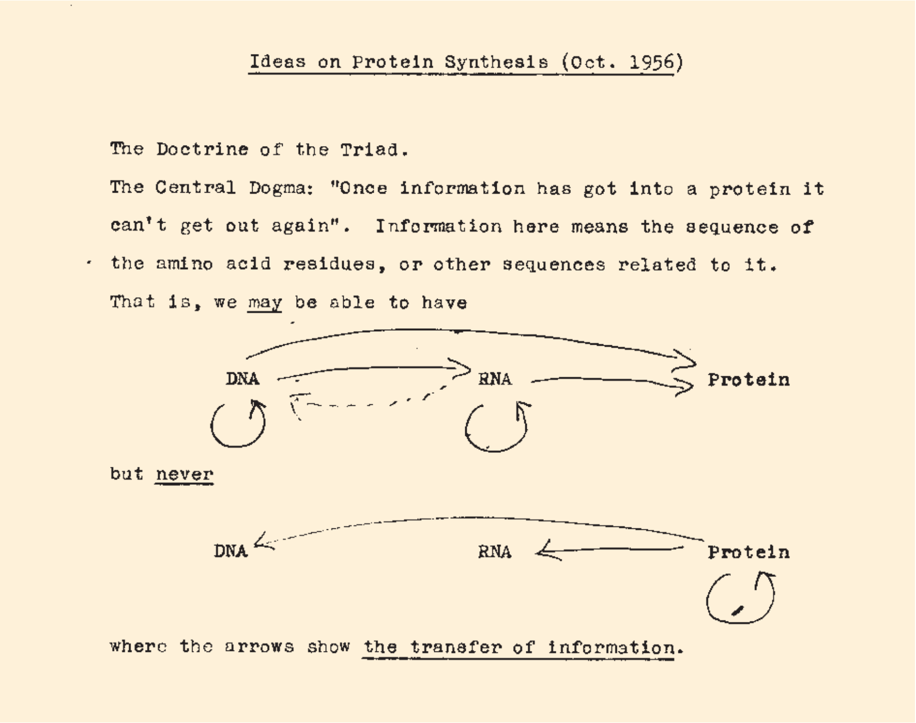
By F. H. C. CRICK

Medical Research Council Unit for the Study of Molecular Biology,
Cavendish Laboratory, Cambridge

The importance of proteins

It is an essential feature of my argument that in biology proteins are uniquely important. They are not to be classed with polysaccharides, for example, which by comparison play a very minor role. Their nearest rivals are the nucleic acids. Watson said to me, a few years ago, 'The most significant thing about the nucleic acids is that we don't know what they do.' By contrast the most significant thing about proteins is that they can do almost anything. In animals proteins are used for structural purposes, but this is not their main role, and indeed in plants this job is usually done by polysaccharides. *The main function of proteins is to act as enzymes.* Almost all chemical reactions in living systems are catalysed by enzymes, and all known enzymes are proteins. It is at first sight paradoxical that it is probably easier for an organism to produce a new protein than to produce a new small molecule, since to produce a new small molecule one or more new proteins will be required in any case to catalyse the reactions.

I shall also argue that the main function of the genetic material is to control (not necessarily directly) the synthesis of proteins. There is a little



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

DNA
RNA
Protein

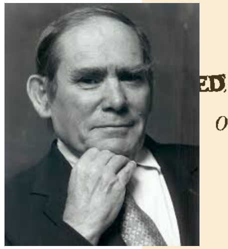
but never

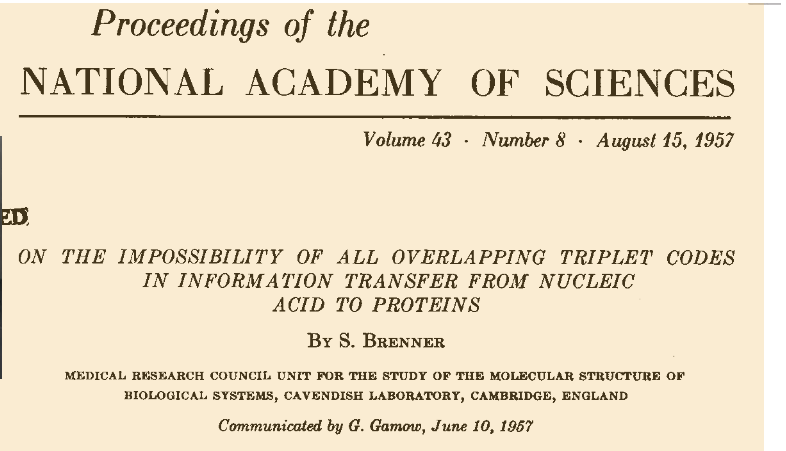
DNA
RNA
Protein

where the arrows show the transfer of information.

FH. Crick. *Symp Soc. Exp. Biol.* 12:138-163 (1958)

Discovery of the Genetic Code for Proteins





Proceedings of the

NATIONAL ACADEMY OF SCIENCES

Volume 43 · Number 8 · August 15, 1957

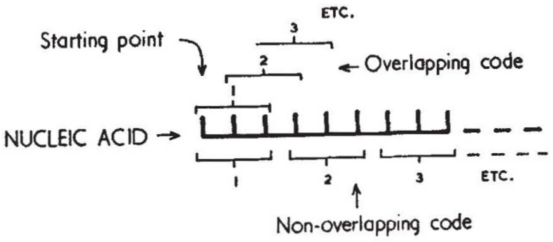
ED

ON THE IMPOSSIBILITY OF ALL OVERLAPPING TRIPLET CODES IN INFORMATION TRANSFER FROM NUCLEIC ACID TO PROTEINS

By S. Brenner

MEDICAL RESEARCH COUNCIL UNIT FOR THE STUDY OF THE MOLECULAR STRUCTURE OF BIOLOGICAL SYSTEMS, CAVENDISH LABORATORY, CAMBRIDGE, ENGLAND

Communicated by G. Gamow, June 10, 1957



- Properties of the general overlapping triplet code:

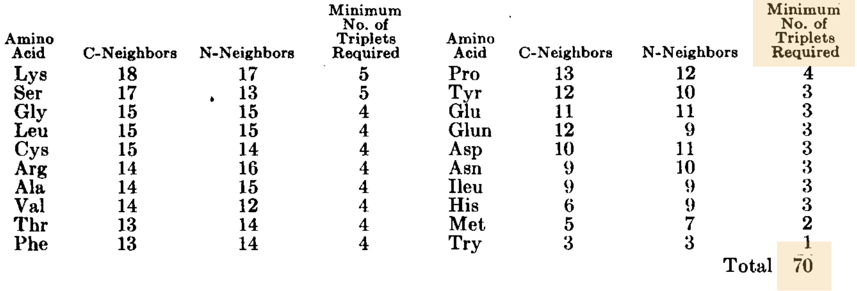
  Coding triplets from 4 nt: maximum of 64 triplets
  Each triplet shares 2 nt with next triplet.
  Degeneracy: 64 triplets degenerated in 20 aa.

- Therefore, peptide sequences would be constrained:

  There could not be more than 256 dipeptide sequences (represented by sequence of 4 nt).
  Yet there are in theory 400 dipeptide sequences.

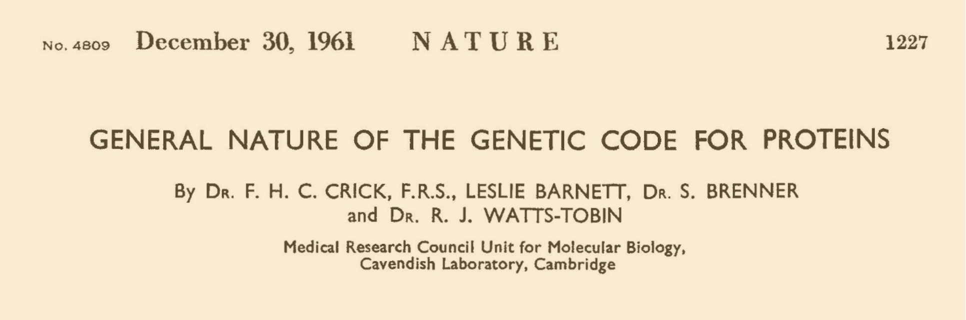
- *Proof* based on data (*reductio ad absurdum*): 64 triplets are insufficient to code the known aa sequences.

  Any triplet can be preceded (or succeeded) by only 4 different nucleotides, hence 4 different triplets.
  Consider *j*,*k*,*l* aa. For every triplet for *k*, there are at most 4 *j* N-neighbour, and 4 *l*, C-neighbours. One can count the minimum of triplets required to encode *k* to account for the largest number of neighbours.



Amino Acid	C-Neighbors	N-Neighbors	Minimum No. of Triplets Required	Amino Acid	C-Neighbors	N-Neighbors	Minimum No. of Triplets Required
Lys	18	17	5	Pro	13	12	4
Ser	17	13	5	Tyr	12	10	3
Gly	15	15	4	Glu	11	11	3
Leu	15	15	4	Glun	12	9	3
Cys	15	14	4	Asp	10	11	3
Arg	14	16	4	Asn	9	10	3
Ala	14	15	4	Ileu	9	9	3
Val	14	12	4	His	6	9	3
Thr	13	14	4	Met	5	7	2
Phe	13	14	4	Try	3	3	1
						Total	70

Discovery of the Genetic Code for Proteins

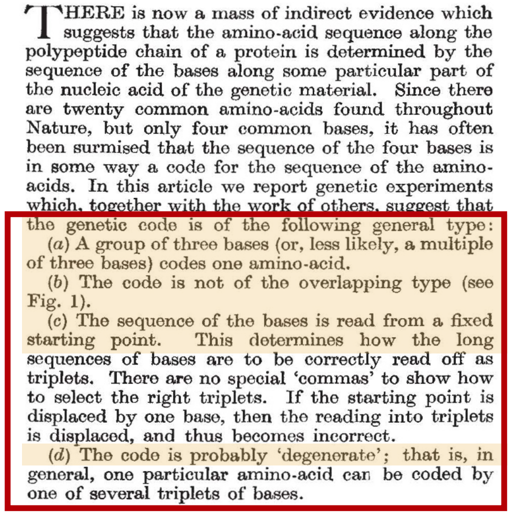


No. 4809 December 30, 1961 NATURE 1227

GENERAL NATURE OF THE GENETIC CODE FOR PROTEINS

By Dr. F. H. C. CRICK, F.R.S., LESLIE BARNETT, Dr. S. BRENNER
and Dr. R. J. WATTS-TOBIN

Medical Research Council Unit for Molecular Biology,
Cavendish Laboratory, Cambridge



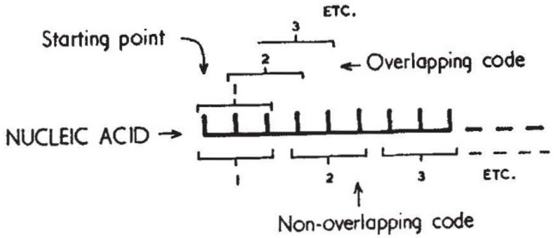
THERE is now a mass of indirect evidence which suggests that the amino-acid sequence along the polypeptide chain of a protein is determined by the sequence of the bases along some particular part of the nucleic acid of the genetic material. Since there are twenty common amino-acids found throughout Nature, but only four common bases, it has often been surmised that the sequence of the four bases is in some way a code for the sequence of the amino-acids. In this article we report genetic experiments which, together with the work of others, suggest that the genetic code is of the following general type:

(a) A group of three bases (or, less likely, a multiple of three bases) codes one amino-acid.

(b) The code is not of the overlapping type (see Fig. 1).

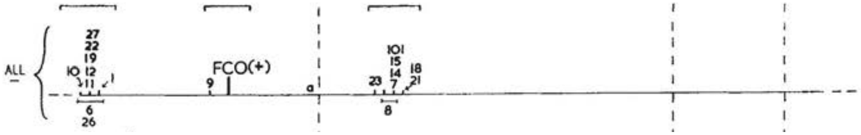
(c) The sequence of the bases is read from a fixed starting point. This determines how the long sequences of bases are to be correctly read off as triplets. There are no special 'commas' to show how to select the right triplets. If the starting point is displaced by one base, then the reading into triplets is displaced, and thus becomes incorrect.

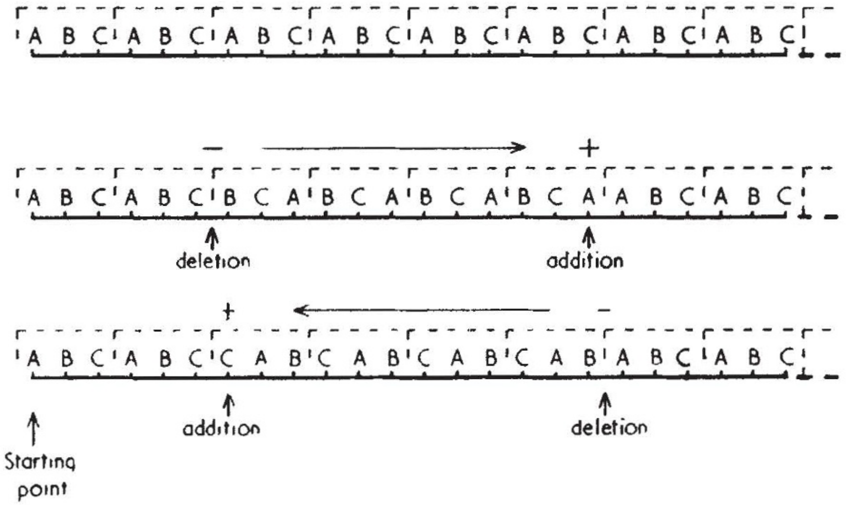
(d) The code is probably 'degenerate'; that is, in general, one particular amino-acid can be coded by one of several triplets of bases.



B cistron of rII region of Bacteriophage T4

(mutants in B will not grow on strain *K12* of *E. coli*)

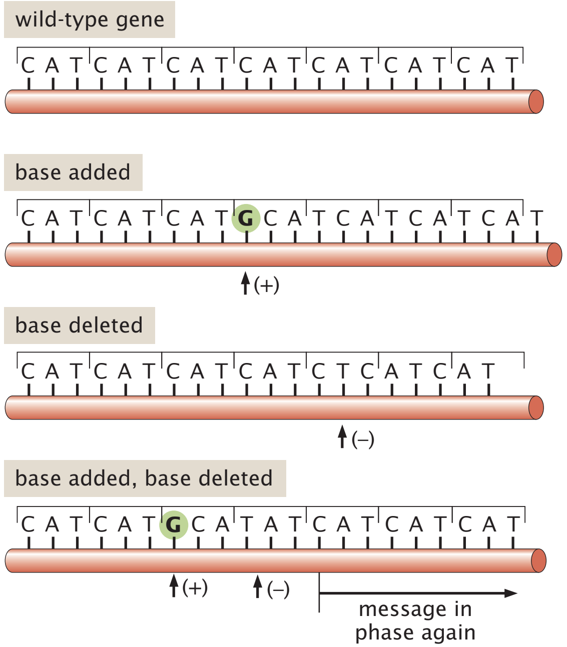



A B C A B C A B C A B C A B C A B C A B C A B C
–
+
A B C A B C B C A B C A B C A B C A A B C A B C
deletion
addition
+
–
A B C A B C C A B C A B C A B C A B A B C A B C
Starting point
addition
deletion

F. Crick, L. Barnett, S. Brenner and J. Watts_Tobin. *Nature* 192: 1227-1232 (1961)

Discovery of the Genetic Code for Proteins

- The code is made of non overlapping triplets of bases


wild-type gene
C A T C A T C A T C A T C A T C A T C A T
base added
C A T C A T C A T G C A T C A T C A T C A T
↑(+)
base deleted
C A T C A T C A T C A T C T C A T C A T
↑(–)
base added, base deleted
C A T C A T G C A T A T C A T C A T C A T
↑(+)
↑(–)
message in phase again

R. Phillips, J. Kondev, J. Thériot & H. Garcia. *Physical Biology of the Cell (Garland Science)* 2012

- Proflavin leads to **addition of base (+); or deletion of base (–)**

both give the *r* phenotype (no plaque, *ie.* non functional T4 on *E. coli K* strain)

- Results:

+ with – reverts to wild type

+ with + or – with – maintain *r* phenotype

+ ; +; + reverts to wild type

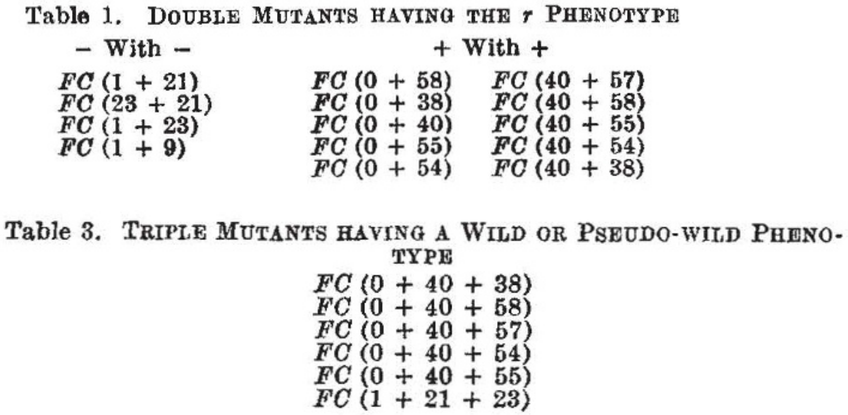


Table 1. DOUBLE MUTANTS HAVING THE *r* PHENOTYPE

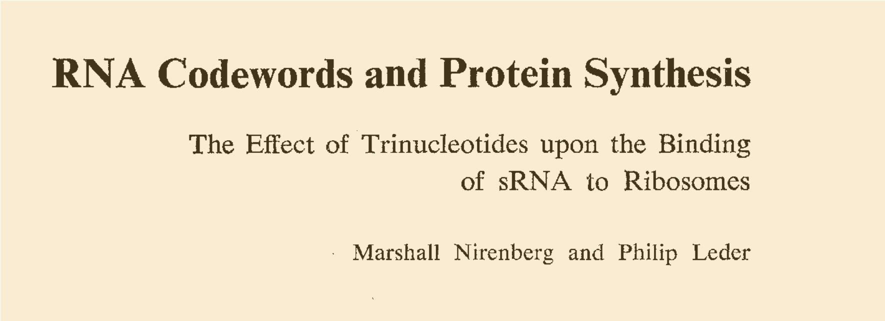
– With –	+ With +	
FC (1 + 21)	*FC* (0 + 58)	*FC* (40 + 57)
FC (23 + 21)	*FC* (0 + 38)	*FC* (40 + 58)
FC (1 + 23)	*FC* (0 + 40)	*FC* (40 + 55)
FC (1 + 9)	*FC* (0 + 55)	*FC* (40 + 54)
	FC (0 + 54)	*FC* (40 + 38)

Table 3. TRIPLE MUTANTS HAVING A WILD OR PSEUDO-WILD PHENOTYPE

FC (0 + 40 + 38)
FC (0 + 40 + 58)
FC (0 + 40 + 57)
FC (0 + 40 + 54)
FC (0 + 40 + 55)
FC (1 + 21 + 23)

F. Crick, L. Barnett, S. Brenner and J. Watts_Tobin. *Nature* 192: 1227-1232 (1961)

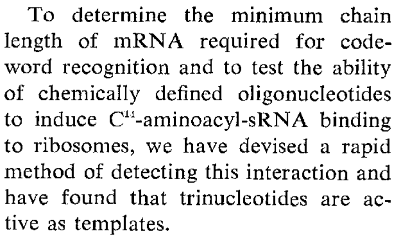
Enraveling the genetic code



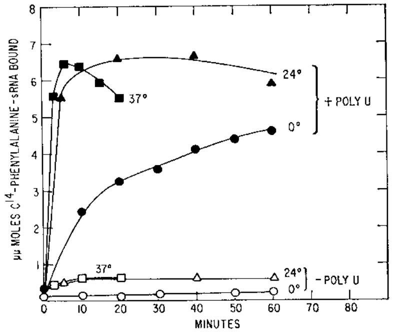
RNA Codewords and Protein Synthesis

The Effect of Trinucleotides upon the Binding of sRNA to Ribosomes

Marshall Nirenberg and Philip Leder



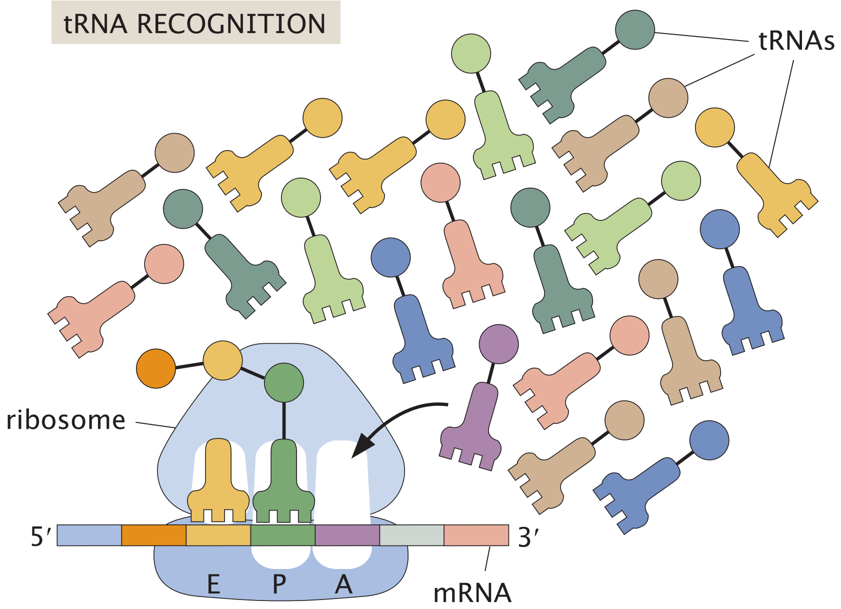
To determine the minimum chain length of mRNA required for codeword recognition and to test the ability of chemically defined oligonucleotides to induce C^{14}-aminoacyl-sRNA binding to ribosomes, we have devised a rapid method of detecting this interaction and have found that trinucleotides are active as templates.



Effect of polyU upon the rate of C^{14}-Phe-transfer tRNA binding to ribosomes

The trinucleotides, pUpUpU, pApApA, and pCpCpC, but not dinucleotides, direct the binding to ribosomes of phenylalanine-, lysine-, and proline-tRNA.

Marshall Nirenberg
(1927-2010)
Nobel 1968



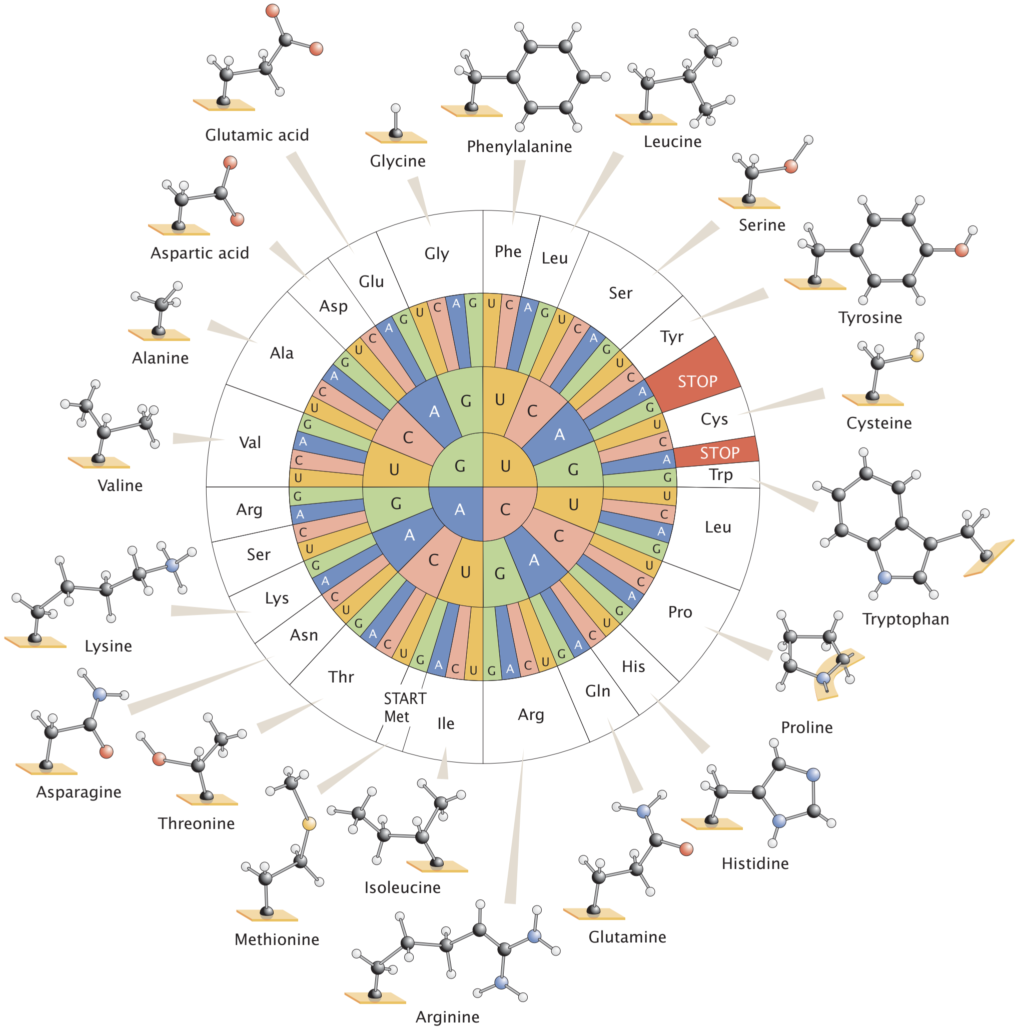
M. Nirenberg and P. Leder. *Science* 145: 1399-1407 (1964)

Enraveling the genetic code

- Facts:
- Triplets of a 4 letter alphabet encode 23 amino acids

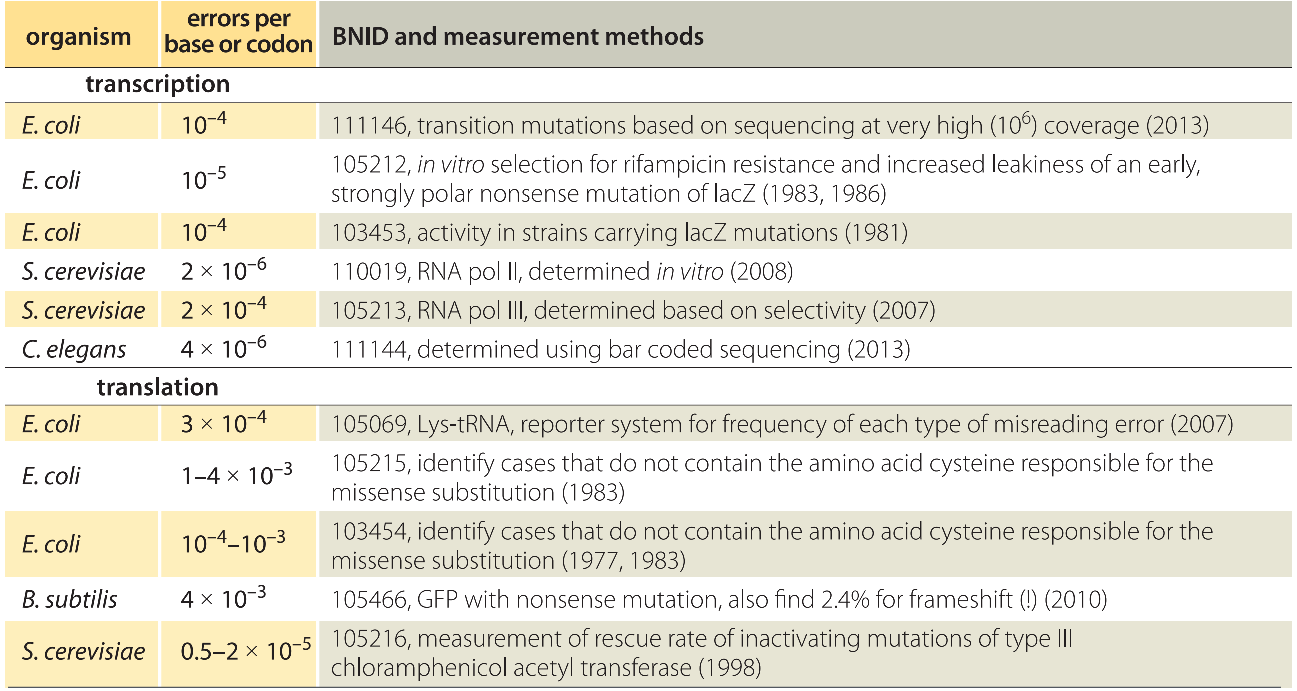
- Properties of the genetic code:
- Degeneracy
- Diversity

- Question: How can a molecular code withstand the impact of noise while accurately and efficiently translating information?



Dealing with errors

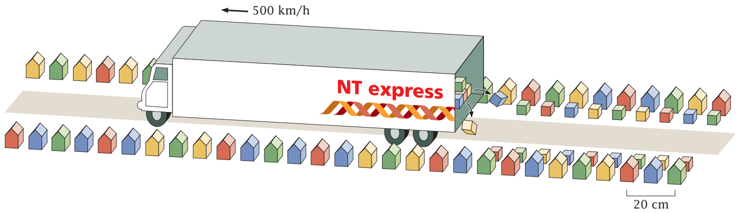
- Measurements of error rates



organism	errors per base or codon	BNID and measurement methods
transcription		
E. coli	10^{-4}	111146, transition mutations based on sequencing at very high (10^6) coverage (2013)
E. coli	10^{-5}	105212, *in vitro* selection for rifampicin resistance and increased leakiness of an early, strongly polar nonsense mutation of lacZ (1983, 1986)
E. coli	10^{-4}	103453, activity in strains carrying lacZ mutations (1981)
S. cerevisiae	2×10^{-6}	110019, RNA pol II, determined *in vitro* (2008)
S. cerevisiae	2×10^{-4}	105213, RNA pol III, determined based on selectivity (2007)
C. elegans	4×10^{-6}	111144, determined using bar coded sequencing (2013)
translation		
E. coli	3×10^{-4}	105069, Lys-tRNA, reporter system for frequency of each type of misreading error (2007)
E. coli	$1–4 \times 10^{-3}$	105215, identify cases that do not contain the amino acid cysteine responsible for the missense substitution (1983)
E. coli	$10^{-4}–10^{-3}$	103454, identify cases that do not contain the amino acid cysteine responsible for the missense substitution (1977, 1983)
B. subtilis	4×10^{-3}	105466, GFP with nonsense mutation, also find 2.4% for frameshift (!) (2010)
S. cerevisiae	$0.5–2 \times 10^{-5}$	105216, measurement of rescue rate of inactivating mutations of type III chloramphenicol acetyl transferase (1998)

R. Milo and R. Phillips Cell Biology by the numbers. *Garland Science*

DNA replication
$10^{-8} – 10^{-10}$

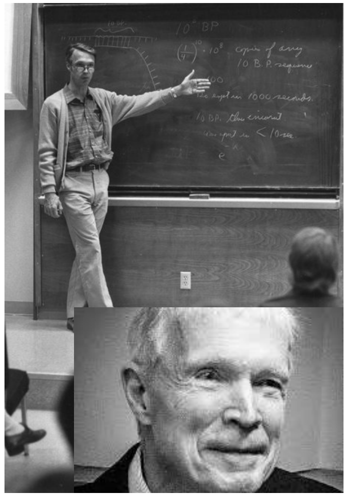


« During replication, the macroscopic replisome travels at a speed of 500 km/h, making a delivery of one of four coloured boxes on both sides of the street every 10 cm, completing its journey (for the case of bacterial replication) in 40 minutes. In this highly efficient delivery process, the truck would deliver a wrong package only once every 3 years! »

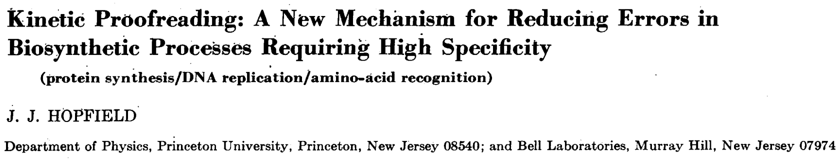
T. Baker
R. Phillips The Molecular Switch. *Princeton Univ. Press*

Dealing with errors

- Kinetic Proofreading



John Hopfield
(1933)

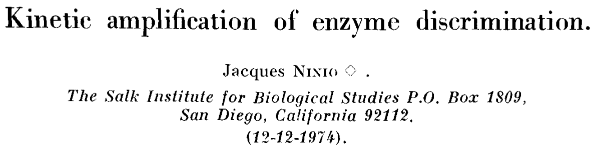


Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity

(protein synthesis/DNA replication/amino-acid recognition)

J. J. HOPFIELD

Department of Physics, Princeton University, Princeton, New Jersey 08540; and Bell Laboratories, Murray Hill, New Jersey 07974



Kinetic amplification of enzyme discrimination.

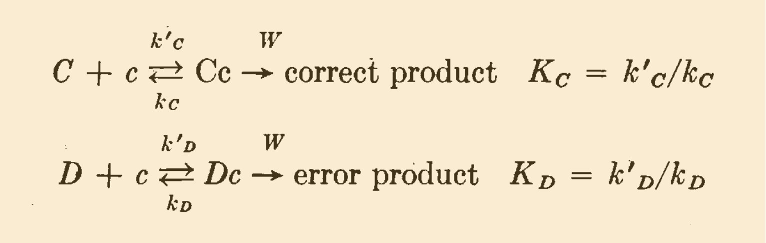
Jacques NINIO ◇ .

The Salk Institute for Biological Studies P.O. Box 1809,
San Diego, California 92112.
(12-12-1974).

J.J. Hopfield (1974) *PNAS* (10): 4135–9

J. Ninio (1975). *Biochimie*. 57 (5): 587–95.

Substrate *C* (resp. *D*) by recognition site *c* (resp.*d):*



$$C + c \underset{k_C}{\overset{k'_C}{\rightleftarrows}} Cc \xrightarrow{W} \text{correct product} \quad K_C = k'_C/k_C$$

$$D + c \underset{k_D}{\overset{k'_D}{\rightleftarrows}} Dc \xrightarrow{W} \text{error product} \quad K_D = k'_D/k_D$$

- <u>Equilibrium discrimination is not sufficient to account for measured error rates.</u>

Error rate = $p_D/p_C = K_C/K_D = e^{-\beta\Delta E_{CD}}$

Energy difference scales with energy associated with formation of 1 H-bond: $\Delta\varepsilon \approx 2k_BT$

Error rate ~e^{-2}~ 0.13

For measured error rate ~10^{-4}

$$\Delta\varepsilon \approx -k_BT \ln 10^{-4} \sim 10 k_BT$$

Dealing with errors

• Kinetic Proofreading

$$C + c \underset{k_C}{\overset{k'_C}{\rightleftarrows}} Cc \xrightarrow{W} \text{correct product} \quad K_C = k'_C/k_C$$

$$D + c \underset{k_D}{\overset{k'_D}{\rightleftarrows}} Dc \xrightarrow{W} \text{error product} \quad K_D = k'_D/k_D$$

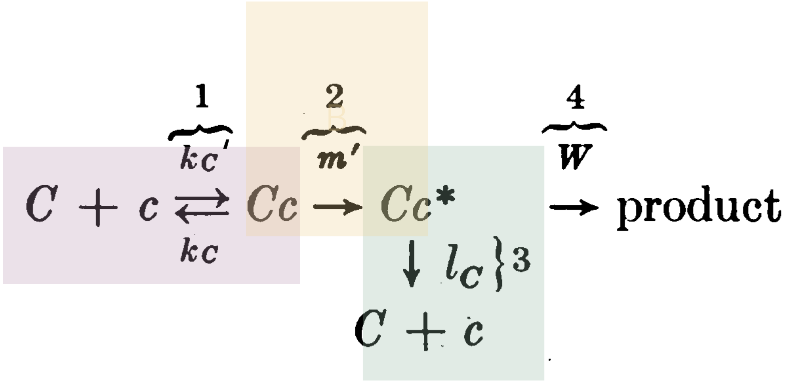
- on rates are the same: $k'_C = k'_D$
  So off rates carry the specificity (ie. discrimination).
- Minimum error attainable :

$$k_C/k_D = K_D/K_C \equiv f_0 = \exp - (\Delta G_{CD}/RT)$$

- Introduction of high energy intermediate *Cc** produced by energy consuming driven reaction (ie. GTP hydrolysis) (2).
- *Cc** dissociates more slowly than *Dc**.
- Two stage kinetic model iterates the same discrimination.

Assuming that m' is substrate independent and that $m' < k_C$ and $W < l_C$, (ie. reactions 1 and 2-3 are at near equilibrium) the error rate is:

$$f = f_{step1} \text{ x } f_{step2} = k_C/k_D \text{ x } l_C/l_D$$



If the same « reading » mechanism is used for dissociation from *Cc** and *Dc** as from *Cc* and *Dc* then: $f = f_0^{\,2}$

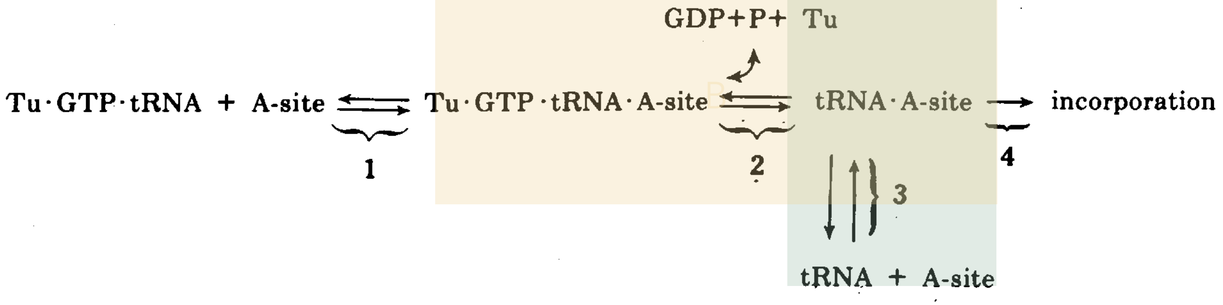
By adding n steps, discrimination is increased with error rate $f = f_0^{\,n+1}$

J.J. Hopfield (1974) *PNAS* (10): 4135–9

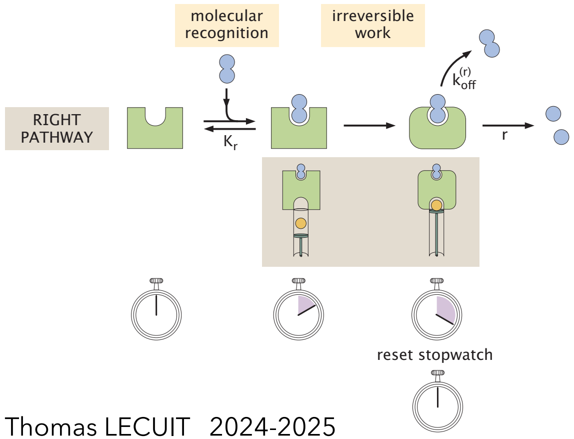
Dealing with errors

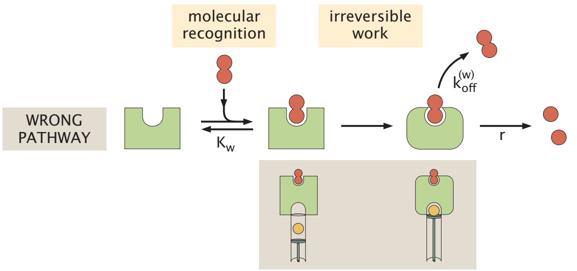
Kinetic Proofreading and translation

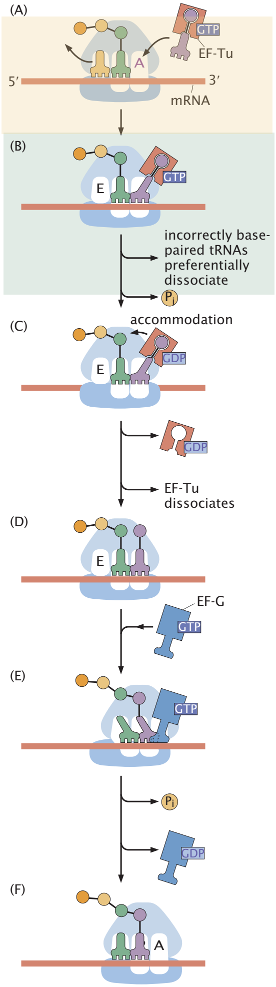
- The aminoacyl-tRNA is activated via GTP hydrolysis before incorporation to the aa chain (by EF-Tu in prokaryotes or eEF1A in eukaryotes).



- **Proofreading can also be interpreted as resulting from the introduction of a lag or delay** (in step 2) that increases reading discrimination between correct and wrong codon/anticodon binding.







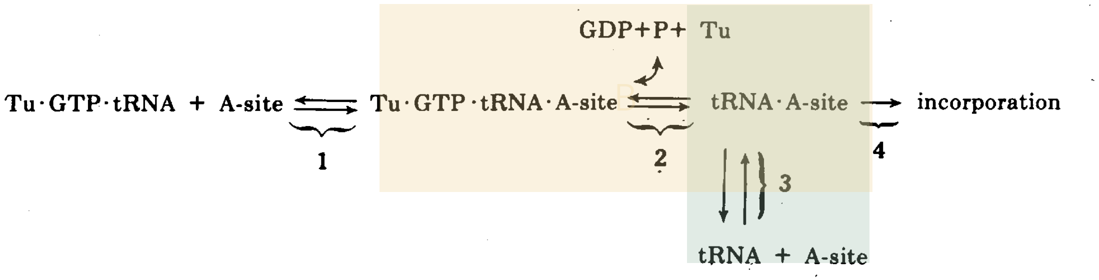
J.J. Hopfield (1974) *PNAS* (10): 4135–9

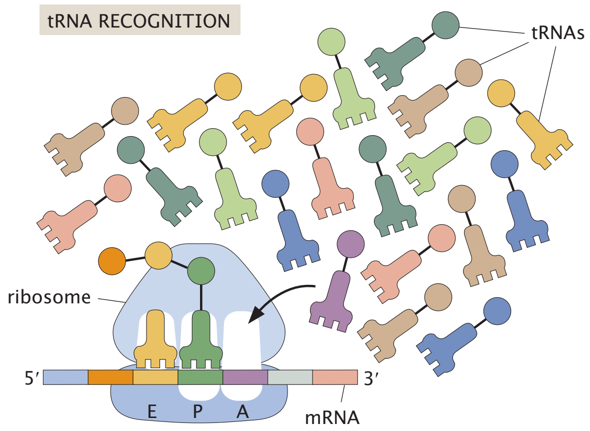
R. Phillips The Molecular Switch. *Princeton Univ. Press*

Dealing with errors

• Kinetic Proofreading and translation

- There is a large excess (few 10 fold) of wrong tRNA-aa competing with a given correct tRNA-aa for binding to an anticodon.





- Error rate in the range of 10^{-4}-10^{-6} /codon should inevitably produce proteins with the wrong amino acid in a cell: the average protein size is ~500aa in mammals, so expected translational error every 2000 average sized protein. There are 10^{10} proteins/cell, so there should be many proteins with wrong incorporated aa. **But this is not the case.**
- **How does the cell cope with this given the impact on function?**
- **How to minimise this error-load?**

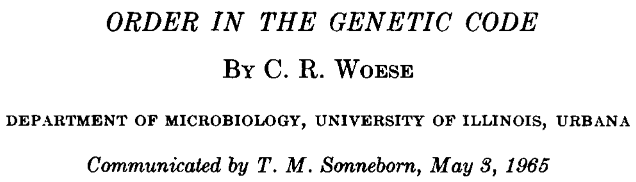
Dealing with errors

• How can the genetic code withstand the impact of noise?

- **Degeneracy of genetic code: many synonymous codons.**

There are potentially 64 different codons. The translation machinery cannot discern well between T and C in 3rd position of codon. Therefore the effective number of codons is at least 48.
Since there are 23 aa, **the code shows degeneracy or redundancy**. All amino acids except methionine and tryptophan are encoded by multiple codons (synonymous codons). Mutations are often synonymous.

- **Carl Woese (1965) – Hypothesis: Close-codons by sequence are either synonymous or encode amino-acids with similar chemical properties.**
- **Smoothness of code table reduces the *error-load* since misreading is likely to replace an amino acid by a chemically related one.**

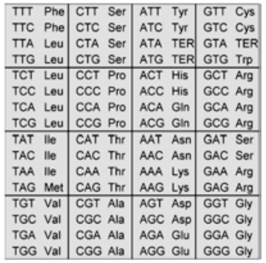

ORDER IN THE GENETIC CODE

BY C. R. WOESE

DEPARTMENT OF MICROBIOLOGY, UNIVERSITY OF ILLINOIS, URBANA

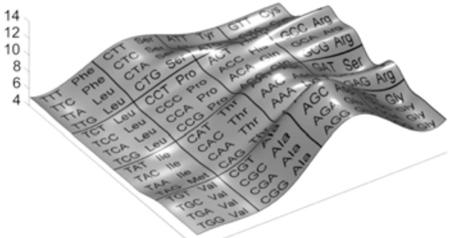
Communicated by T. M. Sonneborn, May 3, 1965

C.R. Woese (1965) *PNAS* (54): 71-75

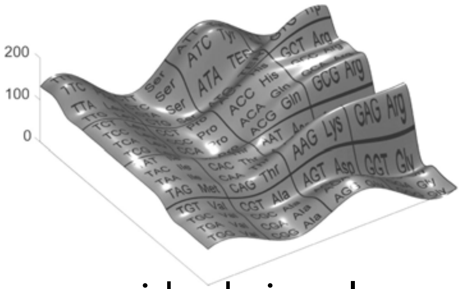


TTT	Phe	CTT	Ser	ATT	Tyr	GTT	Cys
TTC	Phe	CTC	Ser	ATC	Tyr	GTC	Cys
TTA	Leu	CTA	Ser	ATA	TER	GTA	TER
TTG	Leu	CTG	Ser	ATG	TER	GTG	Trp
TCT	Leu	CCT	Pro	ACT	His	GCT	Arg
TCC	Leu	CCC	Pro	ACC	His	GCC	Arg
TCA	Leu	CCA	Pro	ACA	Gln	GCA	Arg
TCG	Leu	CCG	Pro	ACG	Gln	GCG	Arg
TAT	Ile	CAT	Thr	AAT	Asn	GAT	Ser
TAC	Ile	CAC	Thr	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

code table



polarity



side chain volume

T. Tlusty, *Physics of Life Reviews* 7 (2010) 362–376

Evolution of the genetic code

- Hypothesis: The pattern and number of amino-acids are fundamental topological features of the noisy information channel that is embodied in the genetic code. (not a « frozen accident » F. Crick)

- Generic model of genetic code evolution:

T. Tlusty, *Physics of Life Reviews* 7 (2010) 362–376

- Consider 3 features, or forces acting on fitness:
- Diversity: encoding functional proteins requires diverse set of aa. This tends to create a more heterogeneous code.
- Error load: evolutionary selection for codes that minimize the deleterious impact of translation errors and mutations. A mutation should have little impact on chemical nature of translated aa. Error-load selects for smooth code to reduce deleterious effect of mutations.
- Cost of coding system: cost of synthesizing molecules (material, energy and time)

$$\boxed{\mathit{fitness} = -\mathit{error\text{-}load} + w_D \times \mathit{diversity} - w_C \times \mathit{cost}}$$

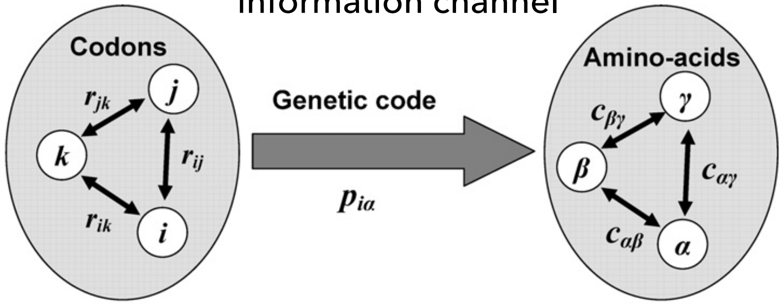
- Coding transition is governed mainly by the cost and quality of this information channel

Evolution of the genetic code

- Coding transition is governed by properties of this information channel

$$fitness = -error\text{-}load + w_D \times diversity - w_C \times cost$$

information channel



N_C codons

N_A amino acids

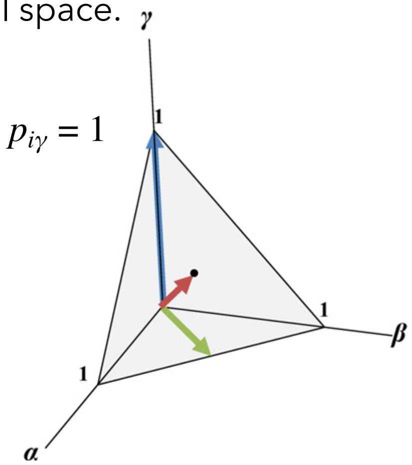
r_{ij}: probability of codon i and j misreading
$c_{\alpha\beta}$: chemical distance

$p_{i\alpha}$ = probability that codon i matches amino acid α
The N_C x N_A probabilities form a **code-matrix.**
Initially the association is random and all $p_{i\alpha} = 1/N_A$ (no information in code): there is no information flow in the channel.

Evolution leads to correlations and information flow.

Code-matrix representation as the ensemble of N_C vectors in an N_A dimensional space.

$$p_{i\alpha} + p_{i\beta} + p_{i\gamma} = 1$$



Blue: codon encodes a single aa
Green: codon encodes 2 aa
Red: codon encodes 3 aa (no information)

- **Smootheness of code** due to error-load tends to **align vectors.**
- **Diversity** tends to bring vectors **in opposite directions.**
- **Cost brings disorder** in vector orientations.

T. Tlusty, *Physics of Life Reviews* 7 (2010) 362–376

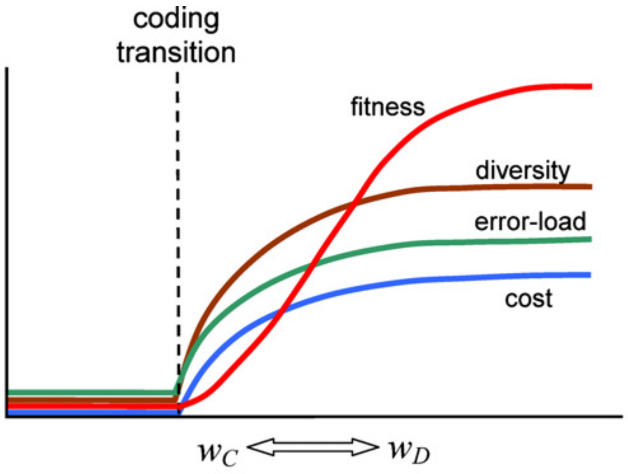
Evolution of genetic code

- Coding transition is governed mainly by properties of this information channel

$$\mathit{fitness} = -\mathit{error\text{-}load} + w_D \times \mathit{diversity} - w_C \times \mathit{cost}$$

- The optimal code is found by maximizing the fitness with respect to the code matrix $p_{i\alpha}$.

- When cost is large (large w_C), specificity is too costly, and the code-matrix is uniform. There is no code.
- When w_C is reduced below a critical value or correspondingly w_D is larger than critical value, certain codon have specificity for aa and there is a coding transition.
- The control parameter in the coding transition is the ratio of w_D and w_C



T. Tlusty, *Physics of Life Reviews* 7 (2010) 362–376

The Genetic Code as a noisy coding system

How to produce a lot from little

- Key feature: Balancing specificity and diversity
  - Increasing diversity imposes a limit on the coding system to ensure specificity
  - The emergence of a smooth code is a solution for dealing with error-load

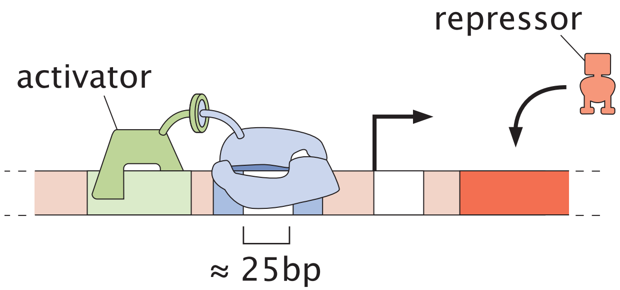
- Balanced by cost

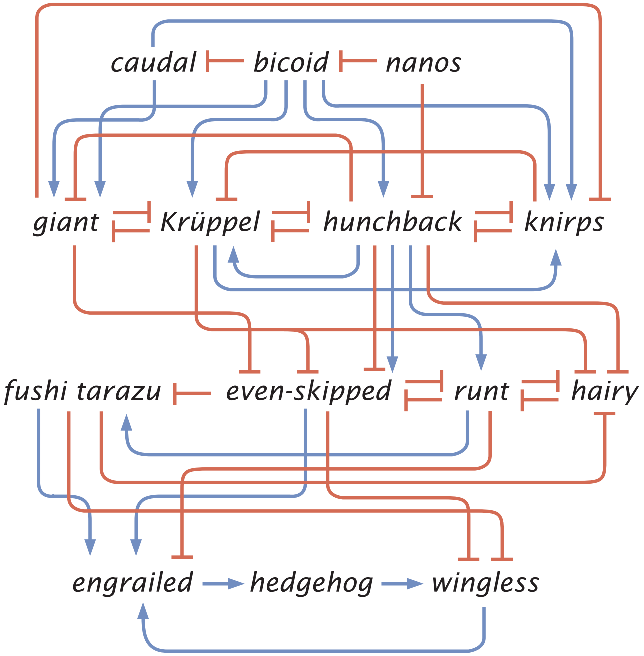
$$\textit{fitness} = -\textit{error-load} + w_D \times \textit{diversity} - w_C \times \textit{cost}$$

Case Study 2: Transcriptional regulatory code

TF/DNA sequences:

how to encode gene expression of ~few 10^4 genes from set of transcription factors (TFs)





Case Study 2: Transcriptional regulatory code

- **Organisms with more genes have more TFs.**
- Proportion of genes coding TFs higher in more complex organisms:
- 169 TFs in yeast and 6275 genes: 2.5%
- 700 TFs in *Drosophila* and 13500 genes: 5%
- 1600 TFs in human, and 20.000 genes: 8%

- Yet: The number of transcription factors within super-families tends to be bounded

- The number of TFs in super-families correlates with degrees of freedom based on number of nucleotide sequences involved in TF binding (few 100s max).

- Organisms with more genes use super-families with larger set of TFs.

Coding limits on the number of transcription factors

Shalev Itzkovitz[1,2], Tsvi Tlusty[2] and Uri Alon*[1,2]

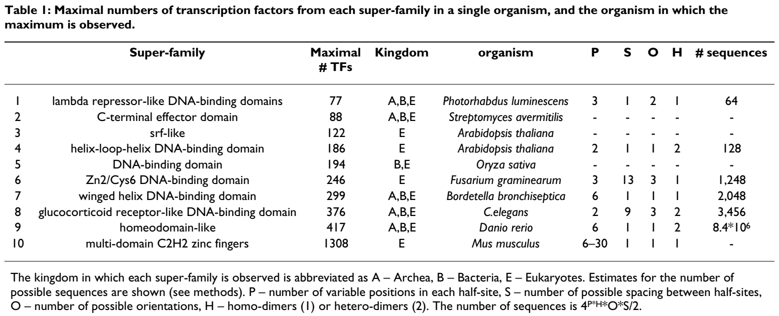
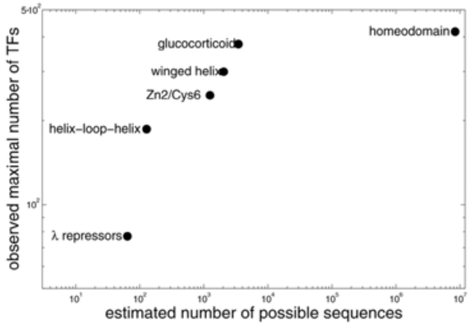
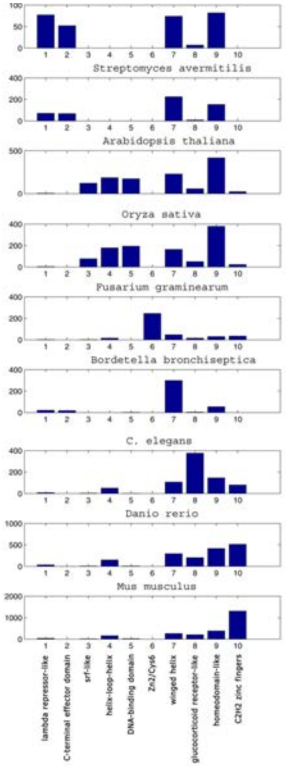


Table 1: Maximal numbers of transcription factors from each super-family in a single organism, and the organism in which the maximum is observed.

	Super-family	Maximal # TFs	Kingdom	organism	P	S	O	H	# sequences
1	lambda repressor-like DNA-binding domains	77	A,B,E	*Photorhabdus luminescens*	3	1	2	1	64
2	C-terminal effector domain	88	A,B,E	*Streptomyces avermitilis*	-	-	-	-	-
3	srf-like	122	E	*Arabidopsis thaliana*	-	-	-	-	-
4	helix-loop-helix DNA-binding domain	186	E	*Arabidopsis thaliana*	2	1	1	2	128
5	DNA-binding domain	194	B,E	*Oryza sativa*	-	-	-	-	-
6	Zn2/Cys6 DNA-binding domain	246	E	*Fusarium graminearum*	3	13	3	1	1,248
7	winged helix DNA-binding domain	299	A,B,E	*Bordetella bronchiseptica*	6	1	1	1	2,048
8	glucocorticoid receptor-like DNA-binding domain	376	A,B,E	*C.elegans*	2	9	3	2	3,456
9	homeodomain-like	417	A,B,E	*Danio rerio*	6	1	1	2	$8.4*10^6$
10	multi-domain C2H2 zinc fingers	1308	E	*Mus musculus*	6–30	1	1	1	-

The kingdom in which each super-family is observed is abbreviated as A – Archea, B – Bacteria, E – Eukaryotes. Estimates for the number of possible sequences are shown (see methods). P – number of variable positions in each half-site, S – number of possible spacing between half-sites, O – number of possible orientations, H – homo-dimers (1) or hetero-dimers (2). The number of sequences is $4^{P*H}*O*S/2$.





S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239

Case Study 2: Transcriptional regulatory code

Model: Protein/DNA as noisy coding system

- **Task:** How to assign different sequences to each transcription factor (TF) in a way that avoids erroneous recognition in which a transcription factor binds the wrong sequence?

- **Complexity and diversity of TFs:** As an organism increases in complexity (eg. cell number, cell types and spatial temporal regulation), there is a need to increase the diversity of gene regulation, via the existence of new TFs.

- **Limit on specificity:** There is a risk that as the # of TFs increases, TFs will become increasingly similar and bind increasingly overlapping sequences. This will limit their specificity.
- **This would thus tend to limit the number of TFs in an organism (similar to amino acids in cells)**

S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239

Case Study 2: Transcriptional regulatory code

Protein/DNA recognition: **a Sphere-packing code.**

- Balancing Diversity and Specificity

The coding problem is akin to a sphere packing code in sequence space.

The size of the sphere of a given TF reflects the extent of binding to *adjacent* sequences

Increasing diversity will lead to overlapping spheres.

This leads to estimates of # of TFs in the same range as observations.

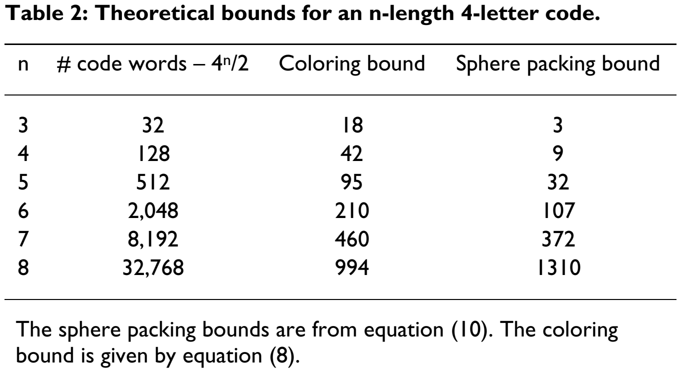
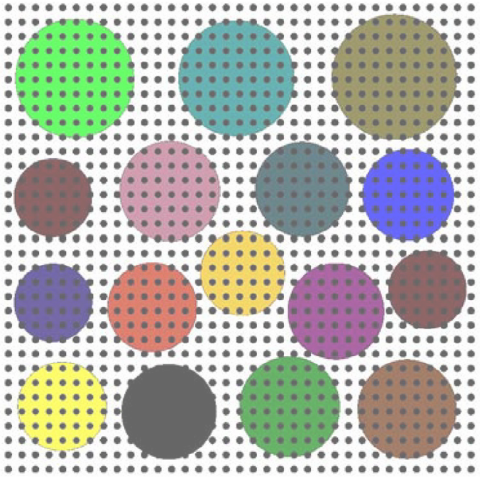


Table 2: Theoretical bounds for an n-length 4-letter code.

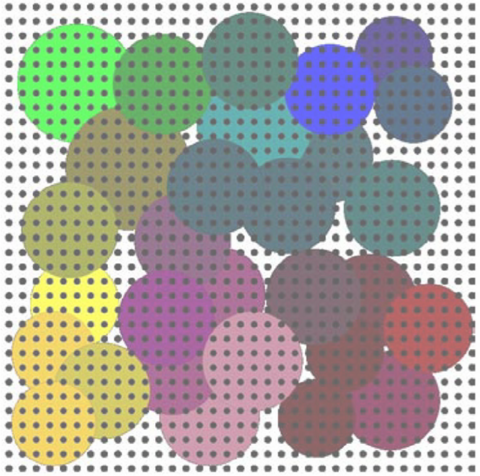
n	# code words – $4^n/2$	Coloring bound	Sphere packing bound
3	32	18	3
4	128	42	9
5	512	95	32
6	2,048	210	107
7	8,192	460	372
8	32,768	994	1310

The sphere packing bounds are from equation (10). The coloring bound is given by equation (8).

Diversity: +/-
Specificity: ++



Diversity: ++
Specificity: +/-



S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239

Case Study 2: Transcriptional regulatory code

Model: Protein/DNA as noisy coding system

- Hypothesis:

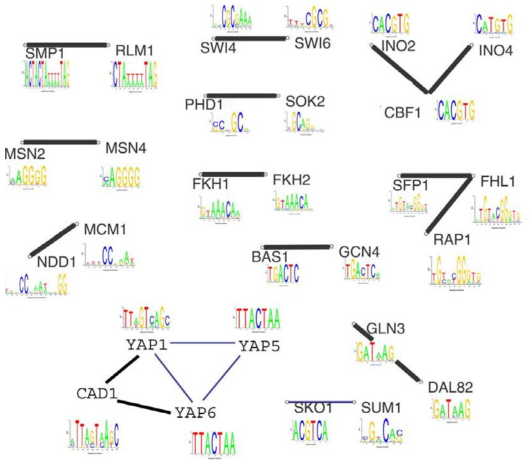
  – Optimal coding theory predicts that overlapping spheres will lead to <u>smooth coding:</u> namely TFs with partially overlapping sequence space should have similar/overlapping functions (regulate overlapping set of genes, to minimise impact on fitness).

- Test:

  – Measure *distance* between TFs based on *sequence bound*

  –Measure *functional distance* (set of regulated genes and/or annotation)

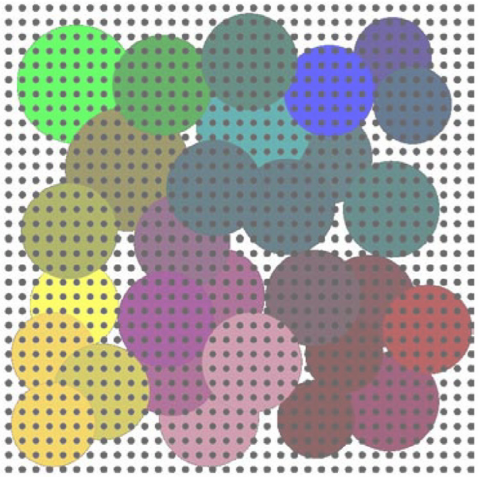
About 14% (276/2016) of all TF pairs had significant target co-regulation. When considering pairs with similar binding sequences, the fraction with significant target co-regulation increases to over 50% (10/18, p-value of $5.1*10^{-5}$)



Transcription factors with overlapping binding sequences in *S. cerevisae*.

Diversity: ++

Specificity: +/-



S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239

Case Study 2: Transcriptional regulatory code

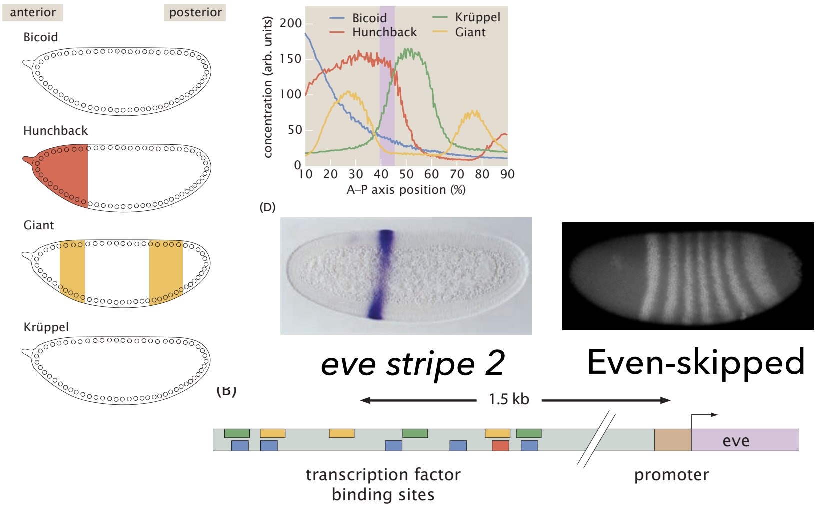
TF/DNA sequences:

how to encode gene expression of ~few 10^4 genes from set of transcription factors (TFs)

Limits imposed on TF diversity in terms of optimal coding strategy.

Yet, there many other layers of regulation of transcription:

- **Combinatorial effects** and integration on regulatory sequences (see course #3).
- Introduces more degrees of freedom and alleviates constraints from sequence overlap

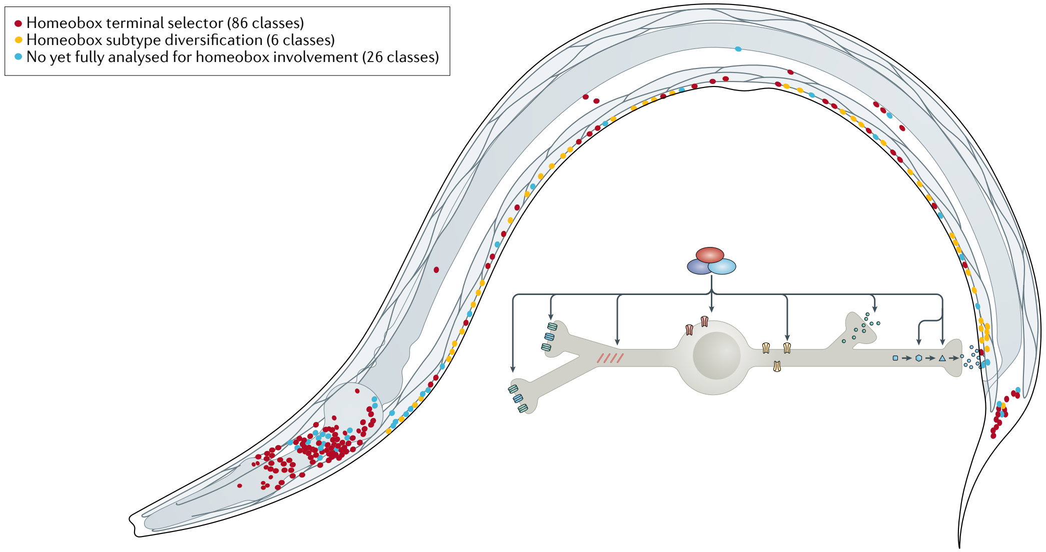


eve stripe 2

Even-skipped

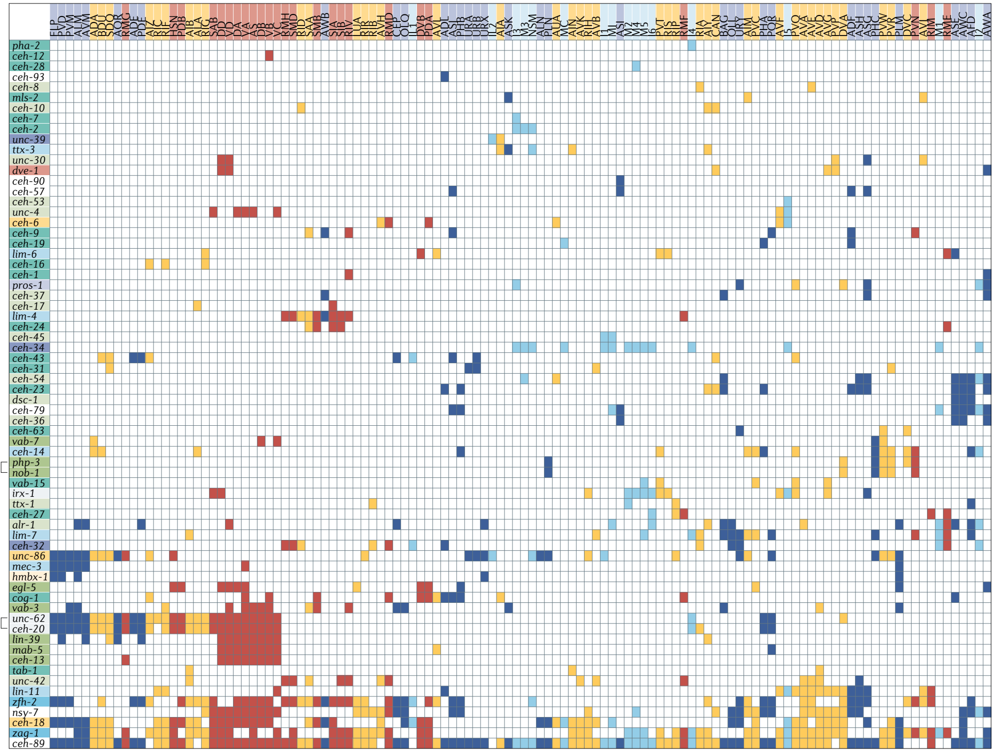
Case Study 2: Transcriptional regulatory code

Combinatorial encoding of cell identity: terminal selectors in nervous system (302 neurons only)


● Homeobox terminal selector (86 classes)
● Homeobox subtype diversification (6 classes)
● No yet fully analysed for homeobox involvement (26 classes)

neuron type

Homeobox gene (TF)



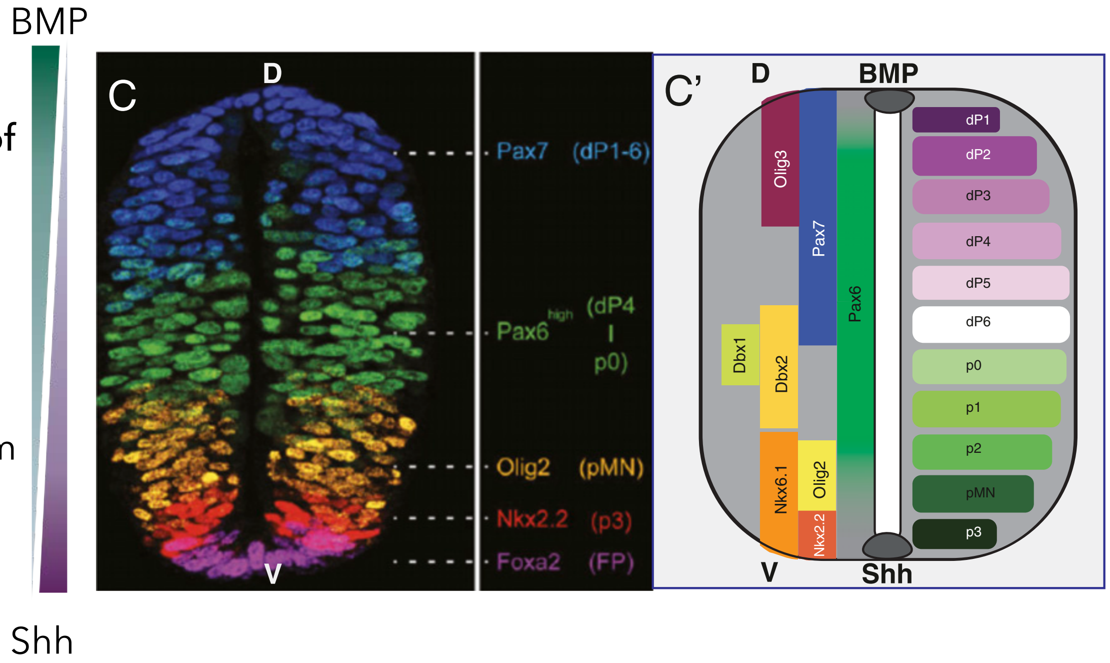
O. Hobert *Nature Neuroscience* 22: 627-636. (2021)

Case Study 2: Transcriptional regulatory code

Spatial combinatorial encoding of cell identity: vertebrate neural tube

- In the developing vertebrate spinal cord, **morphogenetic gradients establish regions of spatially subdivided transcription factor expression that promote cell fate.**

- BMP secreted from the roof plate promotes specification of **dorsal sensory interneurons,** while secretion of Sonic Hedgehog (Shh) from the floor plate specifies **motoneuron/ventral interneuron fate**

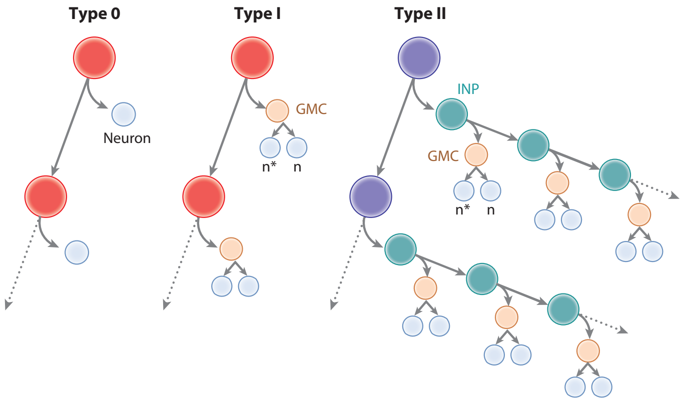


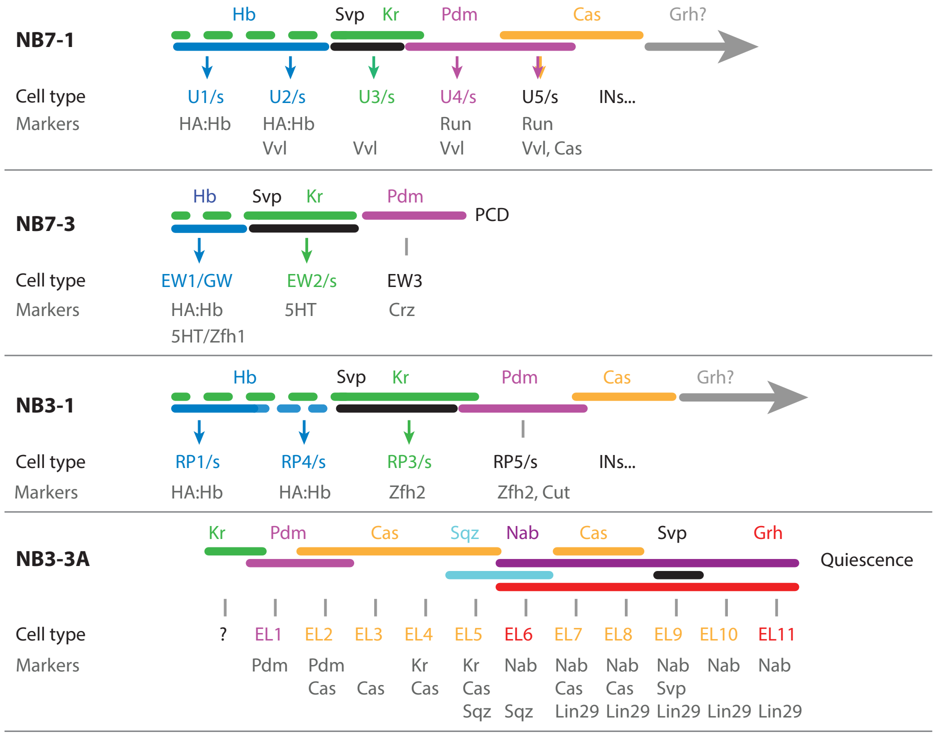
Malin and Desplan C. *PNAS* 2021 Vol. 118 No. 28 e2101823118
(T. Jessel and J. Briscoe labs)

Case Study 2: Transcriptional regulatory code

Temporal combinatorial encoding of cell identity: *Drosophila* nervous system

- In the Ventral Nerve Chain (VNC) distinct neurons are produced in each segment of the embryo
- Progenitor cells divide asymmetrically to generate a new progenitor and a neutron or a GMC which produces 2 neurons or glial cells.
- **A temporal cascade of TFs and spatial cues encode cell identity.**



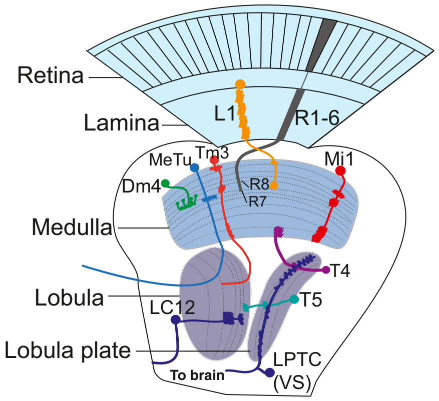


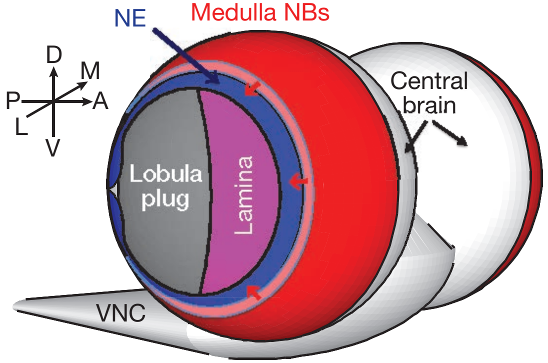
C. Doe *Annu. Rev. Cell Dev. Biol.* 2017. 33:219–40

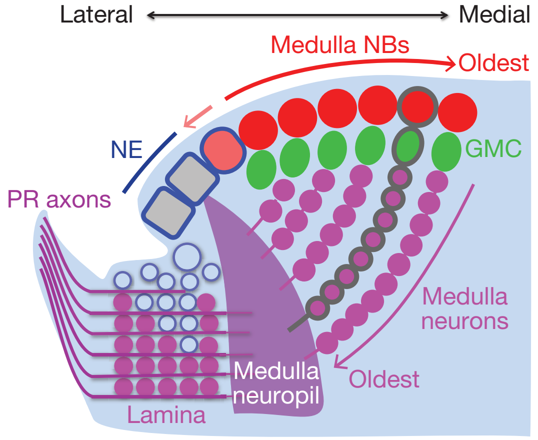
Case Study 2: Transcriptional regulatory code

Temporal combinatorial encoding of cell identity: *Drosophila* nervous system

- 60.000 neurons in the *Drosophila* visual system (⅔ of the brain).
- Visual system requires the specification of distinct neurons.
- 100 Medulla neurons are specified from the Outer Proliferation Center (OPC), in a wave.
- Each progenitor generates a clonal descent of medulla neurons organised in a column.







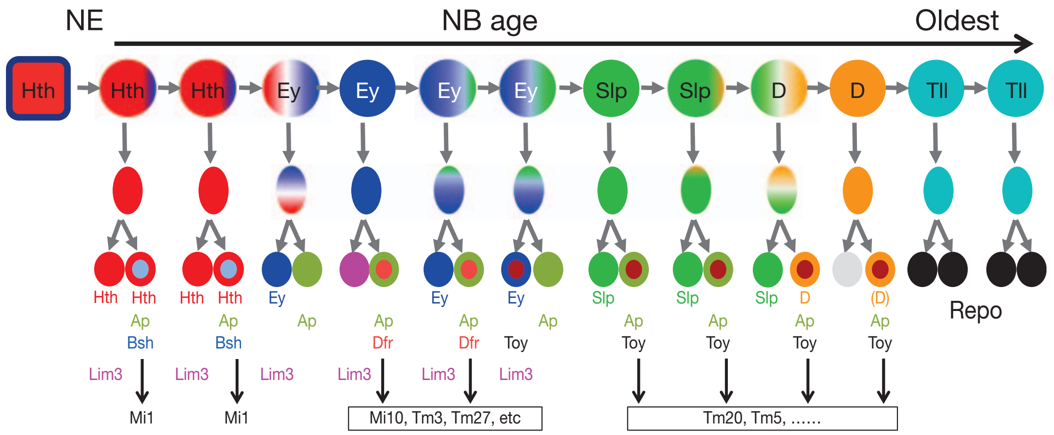
Malin and Desplan C. *PNAS* 2021 Vol. 118 No. 28 e2101823118C.

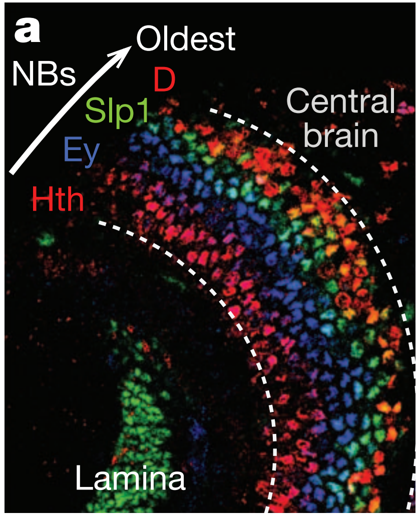
Case Study 2: Transcriptional regulatory code

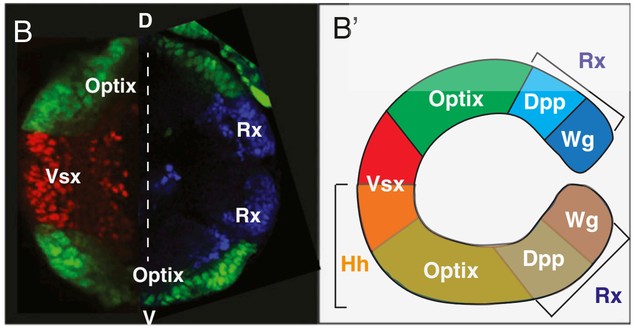
Temporal combinatorial encoding of cell identity: *Drosophila* nervous system

- A temporal cascade to TFs encode distinct cell types.
- This is amplified by spatial cues in the outer proliferation center (OPC).
- Notch signalling amplifies cell diversification (binary choice).

X. Li et al., and C. Desplan. *Nature* 498, 456–462 (2013).



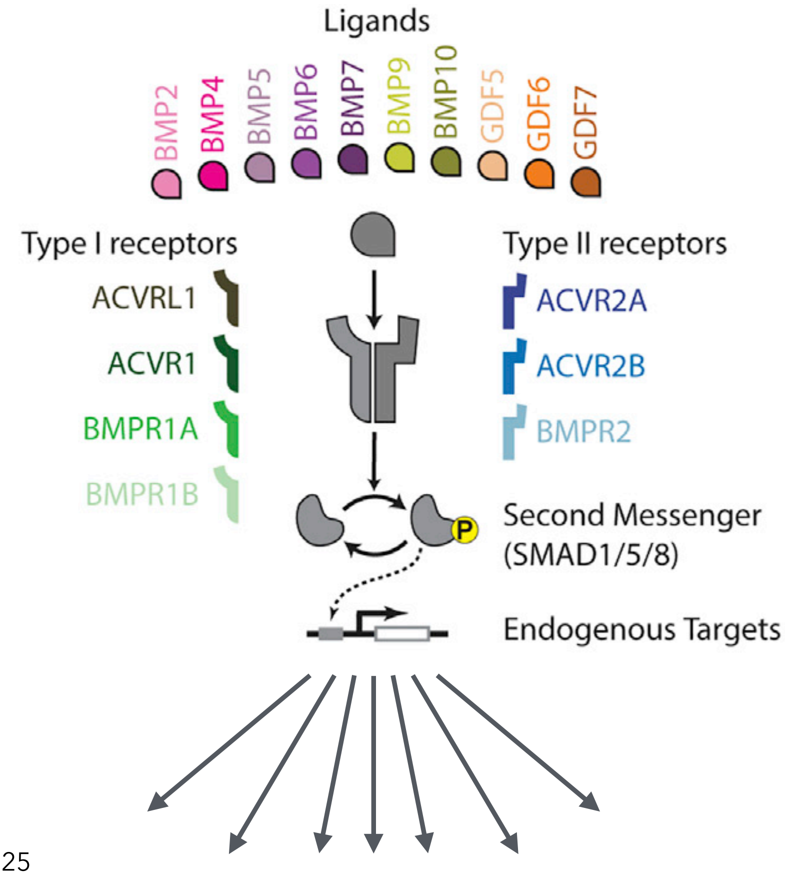




Malin and Desplan C. *PNAS* 2021 Vol. 118 No. 28 e2101823118

Case Study 3: Signal encoding

Ligand/Receptor (signalling code):
how to encode specific signalling output behaviours from different ligands?



Thomas LECUIT 2024-2025

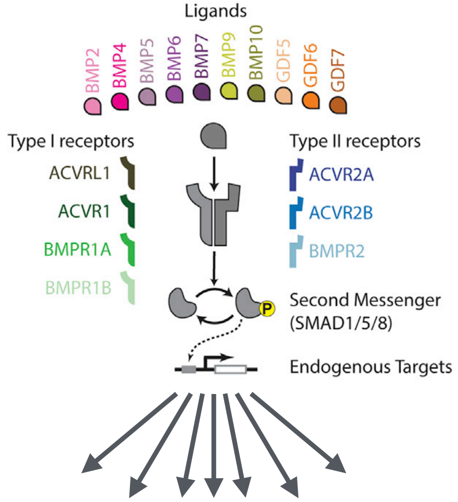
Case Study 3: Signal encoding

There is a **limited set of signalling pathways in metazoans**: 7
BMP, Wnt, Hh, FGF and other RTKs (Receptor Tyrosine Kinase), JNK, TLRs, GPCRs.

- **Ligands/Receptors (L/R) binding trigger specific cellular outputs.**
- **Paradox: Ligands and receptors are generally expressed in poorly restricted domains, yet their function is highly restricted in space and time.**

- **How is specificity achieved? What is the signalling encoding strategy?**

– quantitative control.
– subcellular specification (context dependency)
– ligand diversification



1 to 1 L/R binding
Specific encoding

Promiscuous L/R binding
(many to many)
Combinatorial encoding

>>Power and Limits of these different strategies

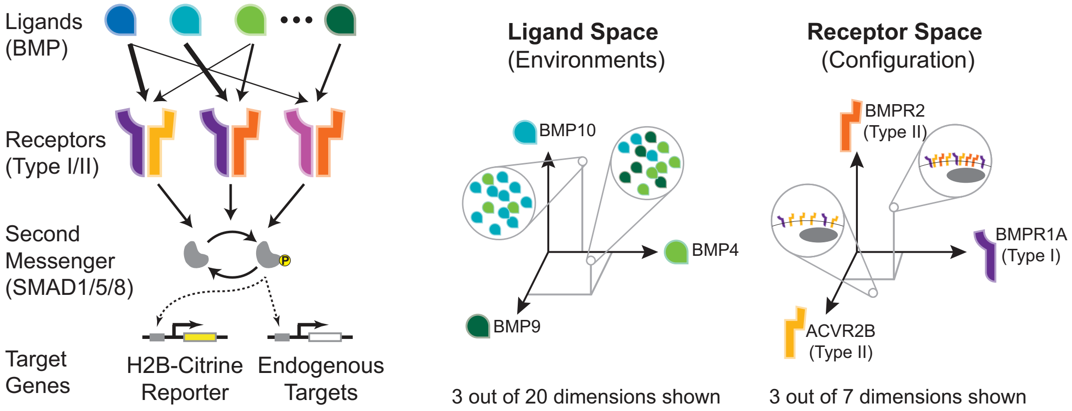
Thomas LECUIT 2024-2025

Case Study 3: Signal encoding

Combinatorial Signal Perception in the BMP Pathway

Yaron E. Antebi,[1] James M. Linton,[1] Heidi Klumpe,[1,2] Bogdan Bintu,[1] Mengsha Gong,[1] Christina Su,[1] Reed McCardell,[1] and Michael B. Elowitz[1,3,4,*]

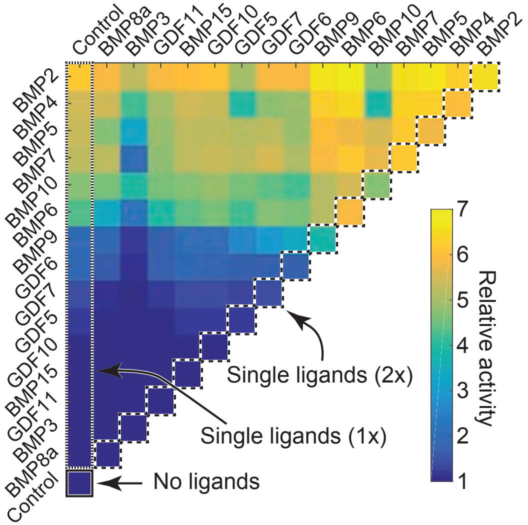
- BMP signalling pathway
- Promiscuous L/R Interactions can be analyzed in terms of multi-dimensional L and R spaces.



3 out of 20 dimensions shown

3 out of 7 dimensions shown

- Combinatorial sensing of BMP ligands.
- Reports BMP signalling with fluorescent reporter using BMP responsive element

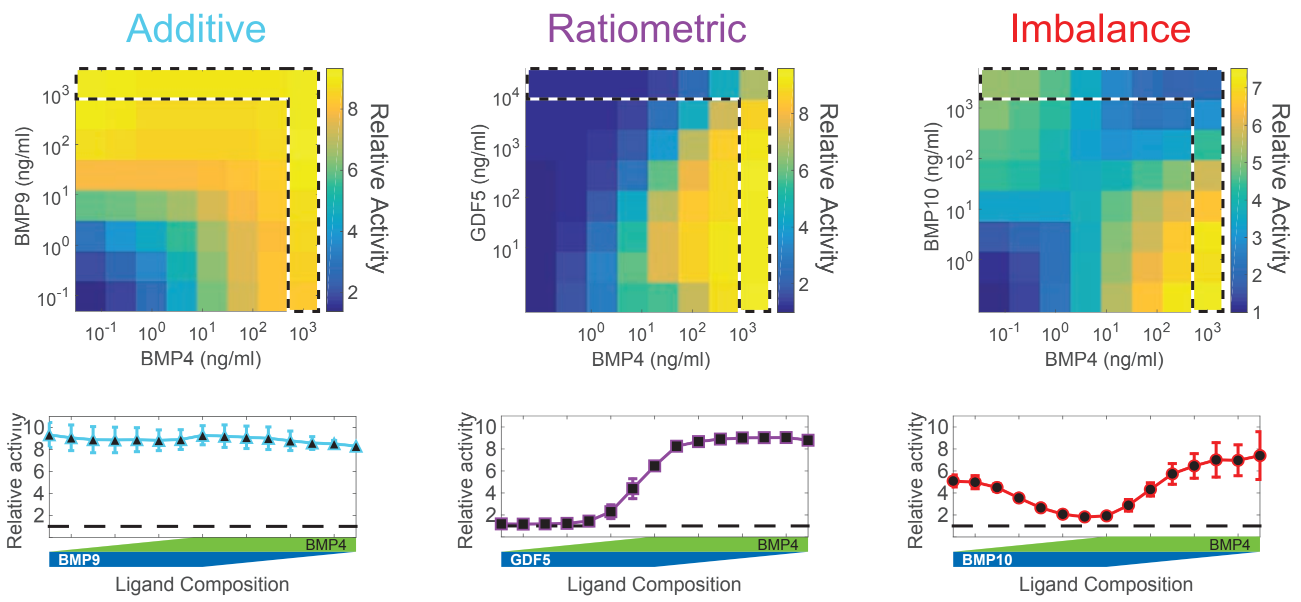


Y.E. Antebi et al. and M.B. Elowitz. *Cell 170*, 1184–1196 (2017)

Case Study 3: Signal encoding

- The combinatorial logic of BMP ligands

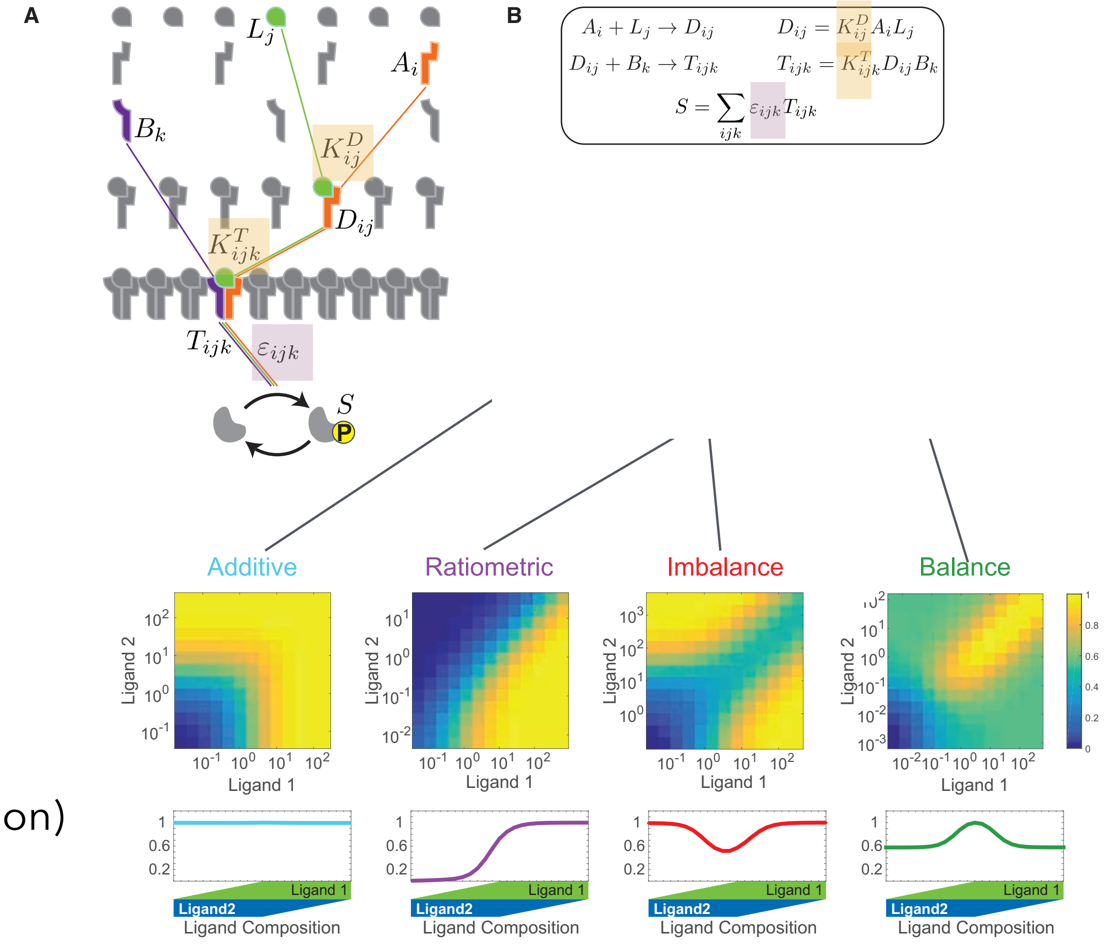
- Binary interaction among BMP ligands reveals different classes of quantitative response behaviours



Y.E. Antebi et al. and M.B. Elowitz. *Cell 170*, 1184–1196 (2017)

Case Study 3: Signal encoding

- **A mathematical models recapitulates the computational properties of BMP signalling**
- Two features are used to assess signalling computation and response profiles across 100.000 parameter sets.
  – **Relative ligand strength to quantify asymmetry in signalling** (activity ratio btw weaker and stronger ligands, equal =1, asymmetry = 0).
  – **Ligand interference coefficient** (+ or - interactions)
- **4 core integration modes or classes of computations emerge**: additive (addition), imbalance (subtraction), balance (multiplication) and ratiometric (division) signalling.



$$A_i + L_j \to D_{ij} \qquad D_{ij} = K^D_{ij} A_i L_j$$

$$D_{ij} + B_k \to T_{ijk} \qquad T_{ijk} = K^T_{ijk} D_{ij} B_k$$

$$S = \sum_{ijk} \varepsilon_{ijk} T_{ijk}$$

Y.E. Antebi et al. and M.B. Elowitz. *Cell 170*, 1184–1196 (2017)

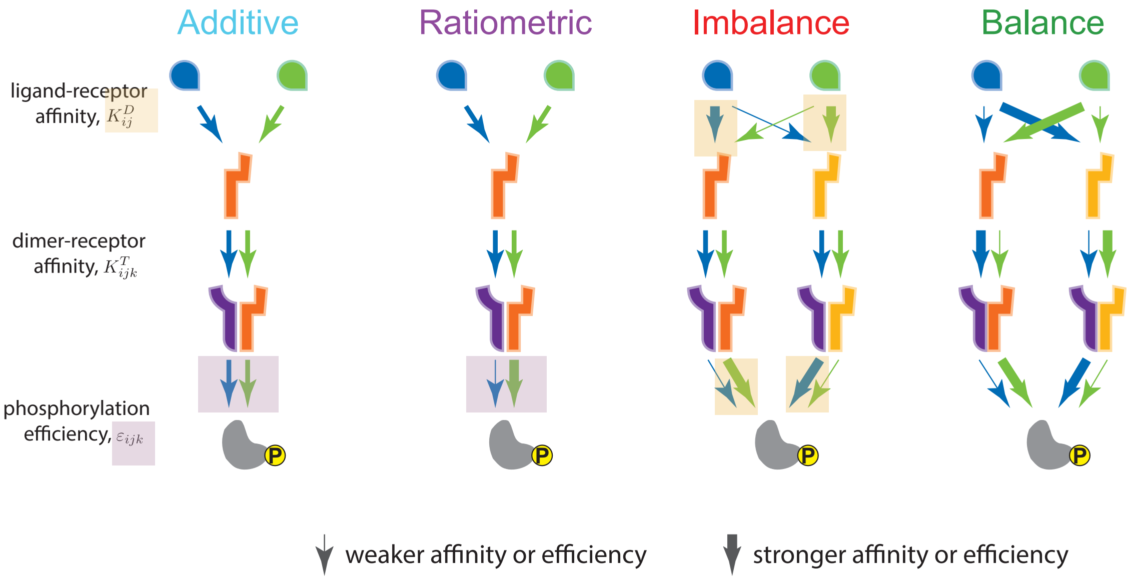
Case Study 3: Signal encoding

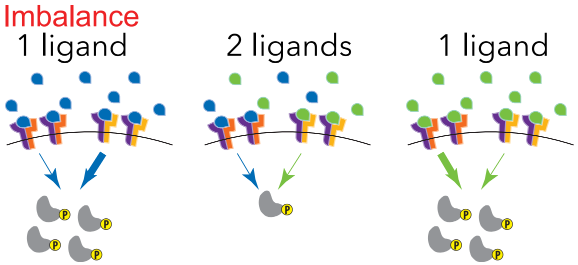
- The 4 computational modes arise through the interplay between different binding affinities (allowing competition for L/R binding) and existence of different complex activities.

- **Additive**: the 2 ligands have ~equivalent activities ($\varepsilon_{i1k} \sim \varepsilon_{i2k}$)
- **Ratiometric**: signaling complexes from one ligand have higher activities than from the other ($\varepsilon_{i1k} \ll \varepsilon_{i2k}$)

- **Imbalance**: each receptor preferentially binds to a distinct ligand with which it forms a less active complex
- **Balance**: each receptor preferentially binds to the ligand with which it forms a more active complex

Y.E. Antebi et al. and M.B. Elowitz. *Cell 170*, 1184–1196 (2017)


Additive
Ratiometric
Imbalance
Balance
ligand-receptor affinity, K_{ij}^D
dimer-receptor affinity, K_{ijk}^T
phosphorylation efficiency, ε_{ijk}
P
weaker affinity or efficiency
stronger affinity or efficiency

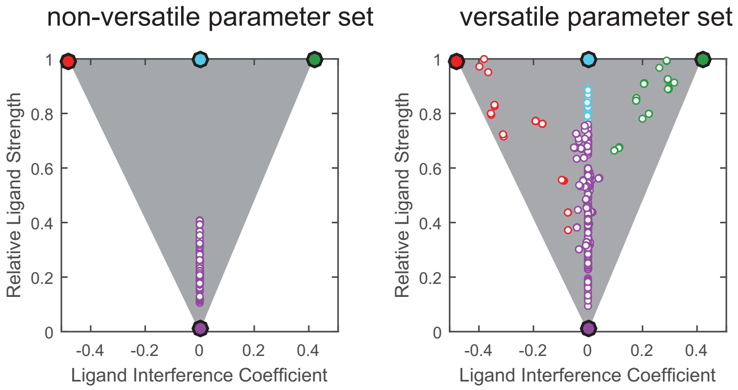


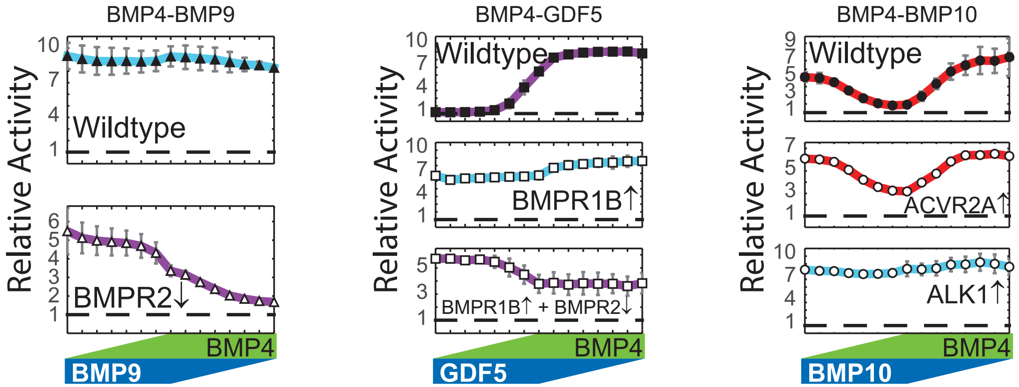
- If only 1 ligand it signals from both receptors
- If 2 ligands, sorting among different receptors and lower total signalling

Case Study 3: Signal encoding

- Different cell types exhibit different computations.
- Receptors are expressed at different levels in different cell types.
- **Receptor expression levels control computation in silico:**
  – change R levels while fixing parameters constant. Some parameters produce few integration modes (left), while others are more versatile and generate different computations (right).

- **Receptor expression levels affect computation in vivo**: additive to radiometric or vice versa, imbalance to additive.

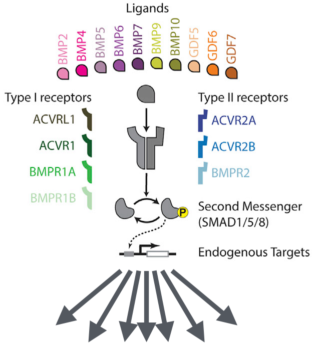




Y.E. Antebi et al. and M.B. Elowitz. *Cell 170*, 1184–1196 (2017)

Case Study 3: Signal encoding

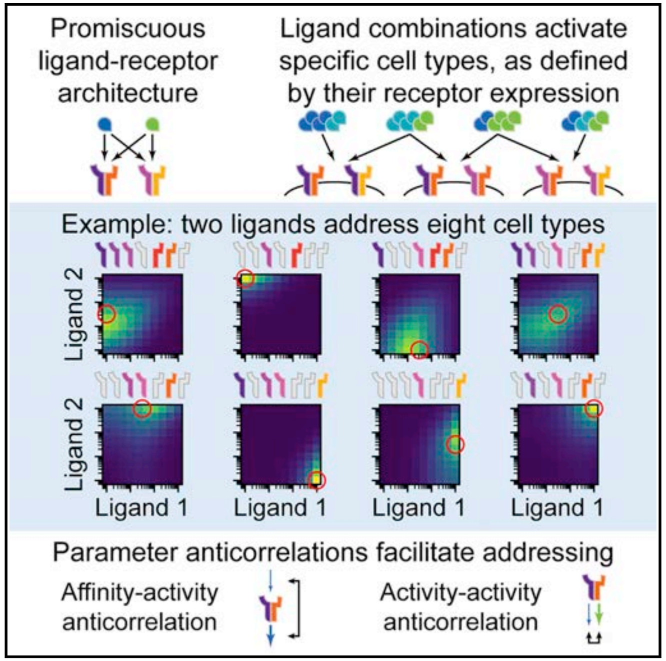
- **Combinatorial signal encoding has following properties:**
- **Sensitivity** to absolute and **relative concentrations**.
- This increases the robustness to variations that affect all ligands in a correlated way (cell surface/cell size or shape, ligand accessibility, etc)



- **Computation is integrated with ligand sensing, and emerges because of decoupling between binding and activity.**
- Computational plasticity: can be tuned eg. by receptor levels.

See also: Su et al., and YE. Antebi, MB. Elowitz *Cell Systems* 13, 408–425 (2022) *Ligand-receptor promiscuity enables cellular addressing*

These interactions allow ligand combinations to selectively activate, or ''address,'' individual cell types or groups of cell types based on their combinatorial receptor expression profiles.

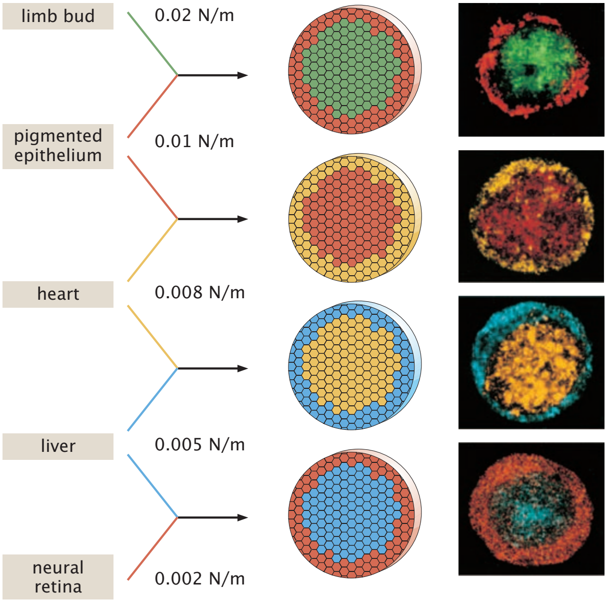


Thomas LECUIT 2024-2025

Case Study 4: Cell-Cell adhesion code

CAM/CAM (adhesion code):

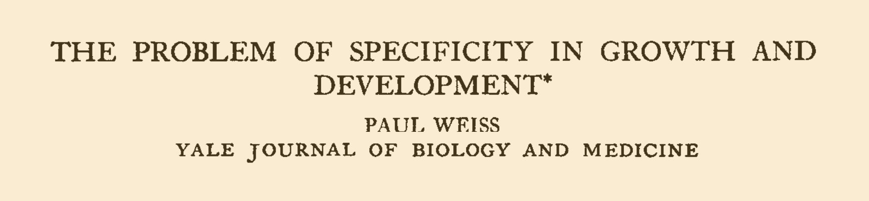
how to encode self-organisation of shapes from few 100 Cell Adhesion Molecules



Case Study 4: Cell-Cell adhesion code

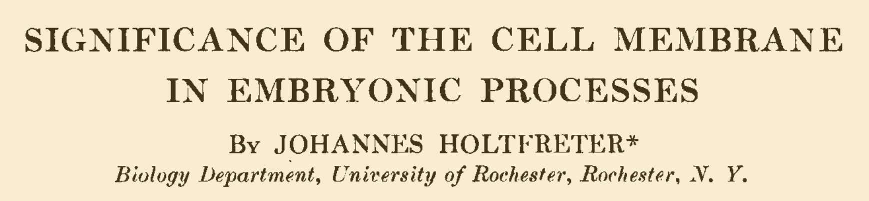
- Interpret cellular affinities in terms of molecular structure and organisation
- Cell surface selective adhesiveness underlies cell clustering during development
  - Specificity: correspondance and mutual fitting between 2 properties

Can be resolved in terms of molecular theory


THE PROBLEM OF SPECIFICITY IN GROWTH AND DEVELOPMENT*

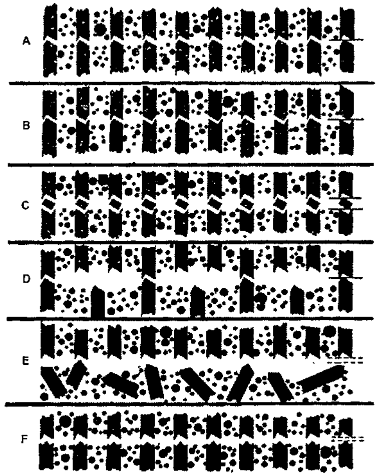
PAUL WEISS

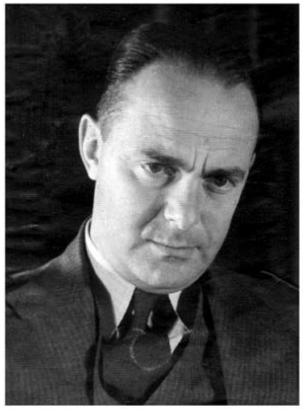
YALE JOURNAL OF BIOLOGY AND MEDICINE


SIGNIFICANCE OF THE CELL MEMBRANE IN EMBRYONIC PROCESSES

By JOHANNES HOLTFRETER*

Biology Department, University of Rochester, Rochester, N. Y.

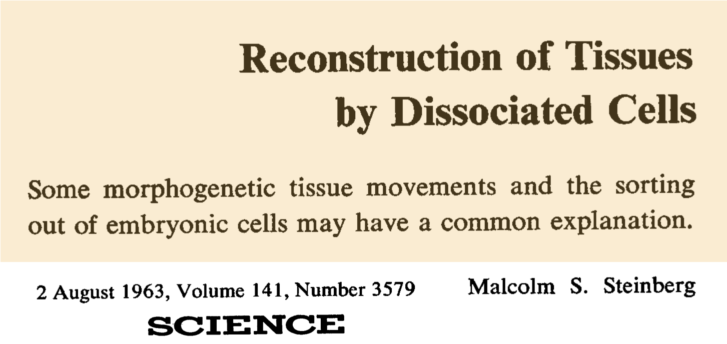



Paul A Weiss (1898-1989)

Johannes Holtfreter 1901- 1992

Case Study 4: Cell-Cell adhesion code

- Encoding tissue organisation via cell-cell adhesion energy



Reconstruction of Tissues by Dissociated Cells

Some morphogenetic tissue movements and the sorting out of embryonic cells may have a common explanation.

2 August 1963, Volume 141, Number 3579 Malcolm S. Steinberg

SCIENCE

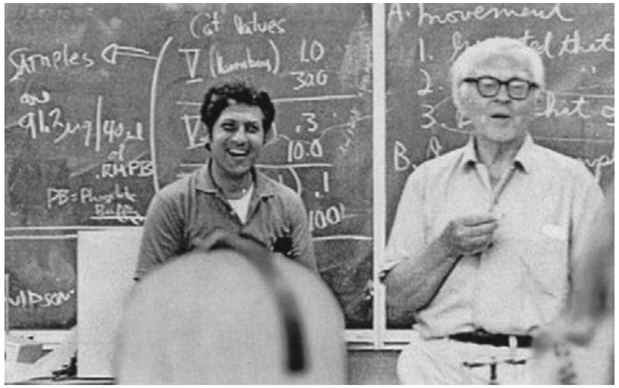
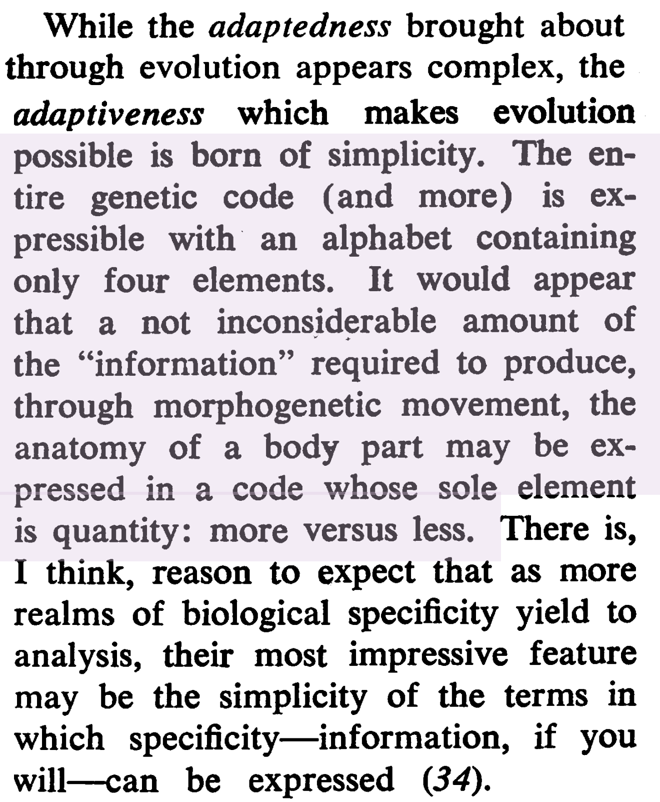


Fig. 2. Hans Holtfreter pronouncing his disagreement with the experimental evidence of Mal Steinberg (shown here laughing) for the thermodynamic model of cell sorting. This photograph was taken at the embryology course at the Marine Biology Laboratory at Woods Hole, 1971.

Steinberg & Gilbert *J. Exp. Zool.* 2004, about Townes & Holtfreter *J. Exp. Zool.* 1955

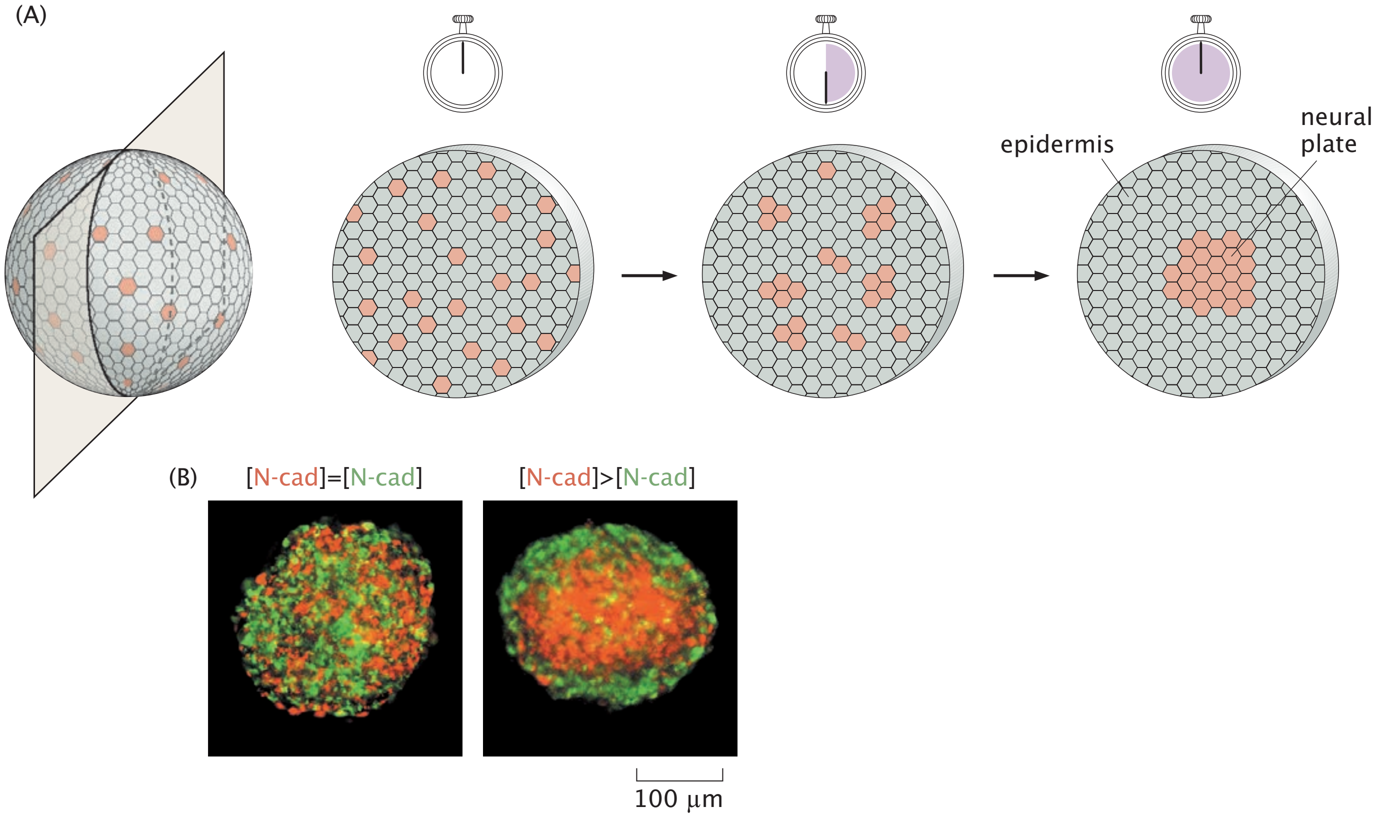


While the *adaptedness* brought about through evolution appears complex, the *adaptiveness* which makes evolution possible is born of simplicity. The entire genetic code (and more) is expressible with an alphabet containing only four elements. It would appear that a not inconsiderable amount of the "information" required to produce, through morphogenetic movement, the anatomy of a body part may be expressed in a code whose sole element is quantity: more versus less. There is, I think, reason to expect that as more realms of biological specificity yield to analysis, their most impressive feature may be the simplicity of the terms in which specificity—information, if you will—can be expressed (*34*).

Steinberg MS. *Science*. 141:401-408. 1963

Case Study 4: Cell-Cell adhesion code

Differential adhesion hypothesis accounts for cell sorting in vitro

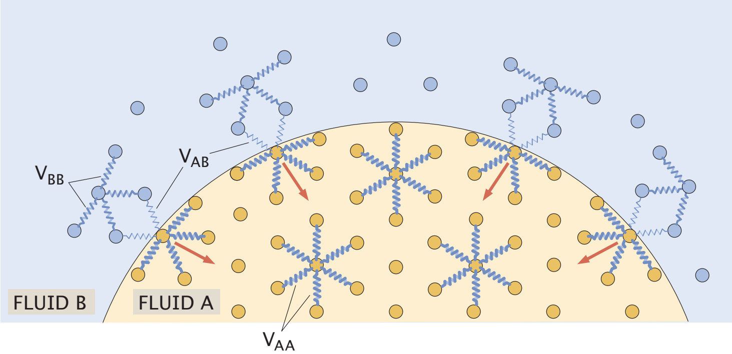


Differential expression levels of a Cadherin drives cell sorting and envelopment behaviour

Based on D. Duguay, R. Foty and MS. Steinberg. *Developmental Biology*. 253: 309-323 (2003)

Case Study 4: Cell-Cell adhesion code

Differential adhesion hypothesis accounts for cell sorting



Interfacial Energy: It is the amount of reversible work to change the surface

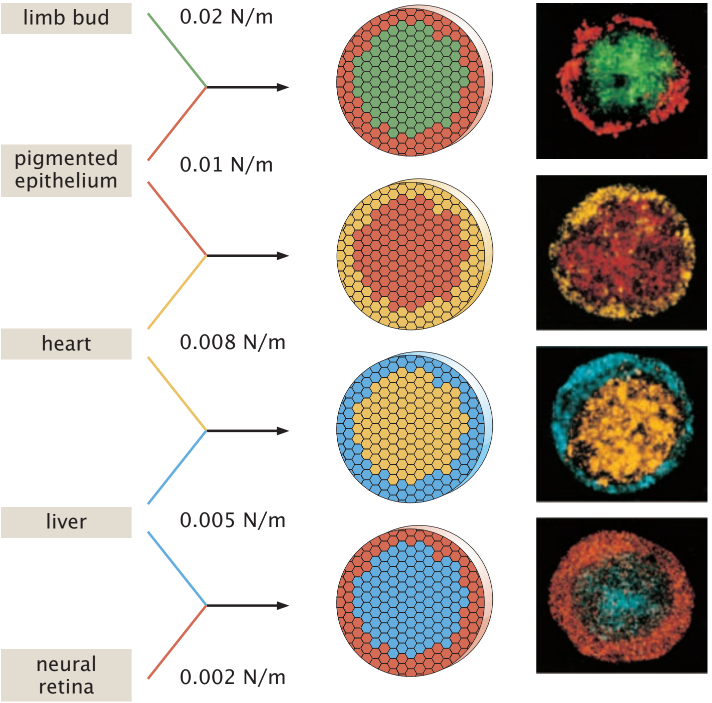
$dE = k\ dS$

Surface tension k (N/m):

– derives from free energy difference between interface and bulk.

– consequence of net inward intermolecular force at interface.

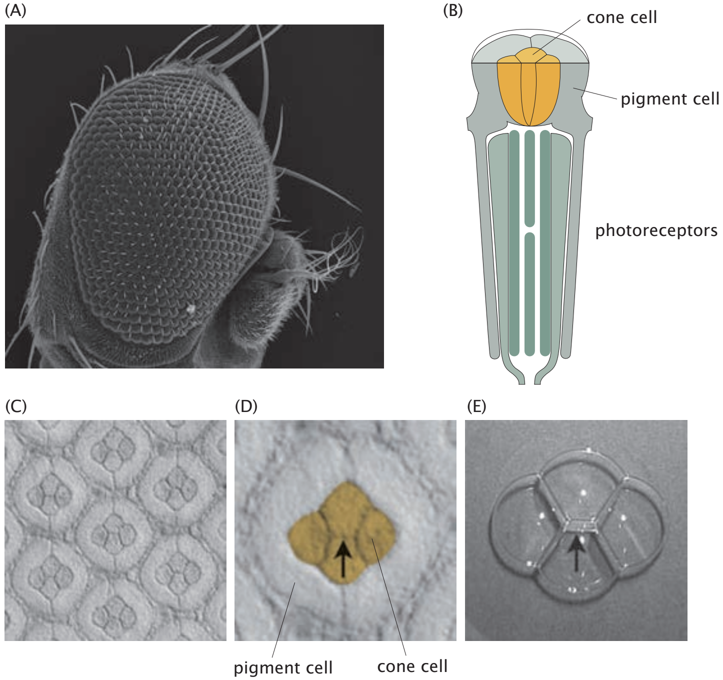
Cell sorting in 3D

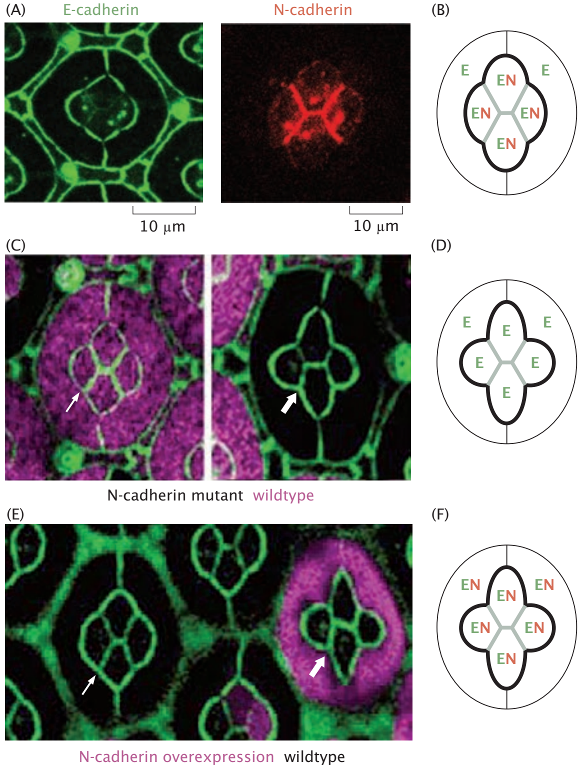


Based on R.Foty et al and MS. Steinberg *Development* 122:1611-1620 (1996)

Case Study 4: Cell-Cell adhesion code

Differential adhesion hypothesis accounts for cell sorting: *Drosophila* retina





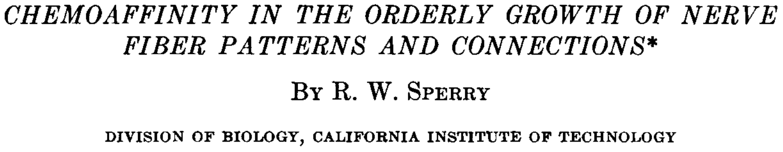
T. Hayashi and R. Carthew. *Nature*. 431:647-652. (2004)

Thomas LECUIT 2024-2025

Case Study 4: Cell-Cell adhesion code

- Chemoaffinity model of nerve routing: Sperry 1963
- Area-code hypothesis: Hood & Dreyer 1977, 1998
- Selective stabilisation by activity: Changeux 1976
- Recognition for synapse specificity?

Reviewed in Sanes JR, Zipursky SL. *Cell*. 181(3):536-556 (2020)



*CHEMOAFFINITY IN THE ORDERLY GROWTH OF NERVE FIBER PATTERNS AND CONNECTIONS**

By R. W. Sperry

DIVISION OF BIOLOGY, CALIFORNIA INSTITUTE OF TECHNOLOGY

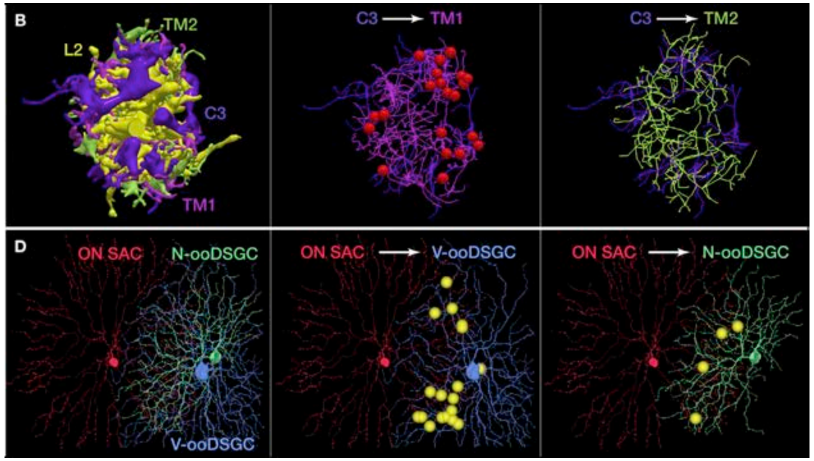
R.W. Sperry. *PNAS* 50(4): 703–710 (1963)

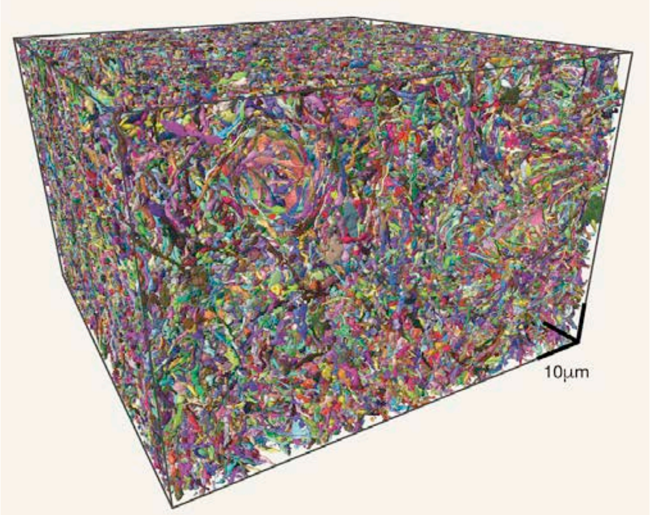
The area code hypothesis revisited: Olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos

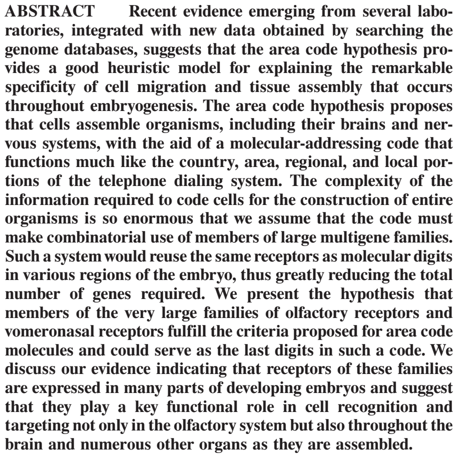
*William J. Dreyer**

Division of Biology, California Institute of Technology, Pasadena, CA 91125

WJ. Dreyer *PNAS* 95:9072–9077 (1998)



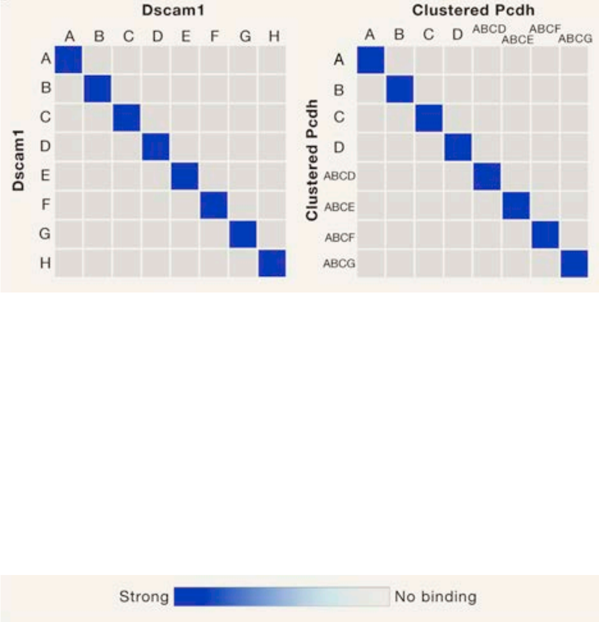




ABSTRACT Recent evidence emerging from several laboratories, integrated with new data obtained by searching the genome databases, suggests that the area code hypothesis provides a good heuristic model for explaining the remarkable specificity of cell migration and tissue assembly that occurs throughout embryogenesis. The area code hypothesis proposes that cells assemble organisms, including their brains and nervous systems, with the aid of a molecular-addressing code that functions much like the country, area, regional, and local portions of the telephone dialing system. The complexity of the information required to code cells for the construction of entire organisms is so enormous that we assume that the code must make combinatorial use of members of large multigene families. Such a system would reuse the same receptors as molecular digits in various regions of the embryo, thus greatly reducing the total number of genes required. We present the hypothesis that members of the very large families of olfactory receptors and vomeronasal receptors fulfill the criteria proposed for area code molecules and could serve as the last digits in such a code. We discuss our evidence indicating that receptors of these families are expressed in many parts of developing embryos and suggest that they play a key functional role in cell recognition and targeting not only in the olfactory system but also throughout the brain and numerous other organs as they are assembled.

Case Study 4: Cell-Cell adhesion code

- Large families with highly specific binding: not a synapse identification code but a code for self/ non-self recognition.
    - Dscam1 (fly): >18.000 splicing isoforms that differ within 3 Ig domain. Isoform specific homophonic recognition. Each neuron expresses 10-40 isoforms in a probabilistic way.
    - Pcdh (vertebrates): 3 tandem genes (58 in total), that multimerize in cis and trans (model: size dependent recognition mechanism)



- Small families with promiscuous binding for subtype recognitions. Ex: DIP/Dpr
- Combinatorial expression. Provides general address code.
- Reuse: to minimise size of families.

Reviewed in Sanes JR, Zipursky SL. *Cell*. 181(3):536-556 (2020)

Case Study 4: Limits of Cell adhesion code

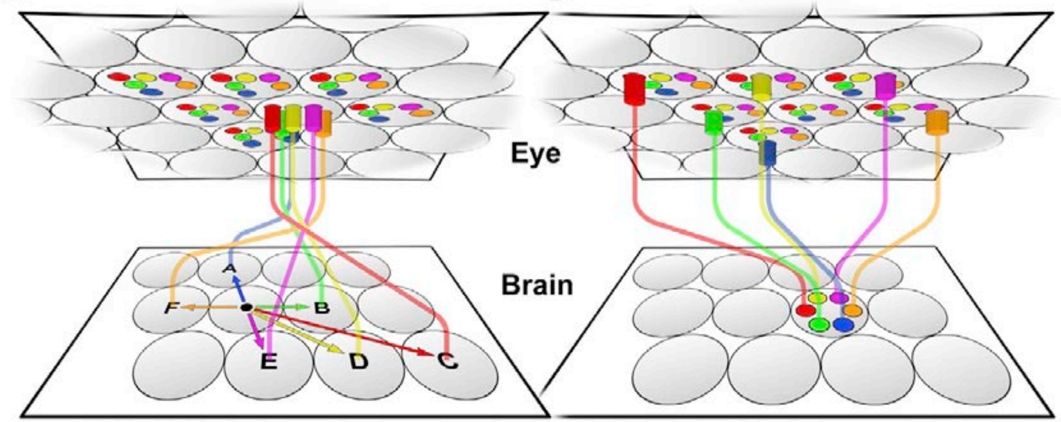
Differential sorting of growth cones without target code recognition

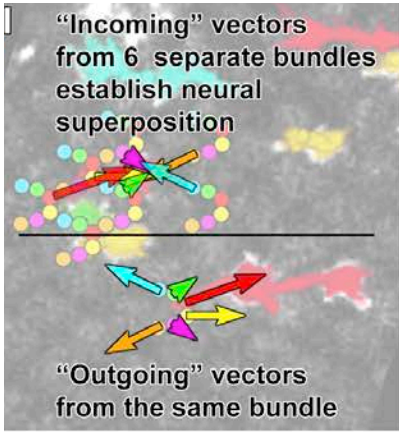
Neural superposition

Axonal self-sorting without target guidance in *Drosophila* visual map formation

Egemen Agi[1]†, Eric T. Reifenstein[2]†, Charlotte Wit[1], Teresa Schneider[1], Monika Kauer[1], Melinda Kehribar[1], Abhishek Kulkarni[1], Max von Kleist[2]*, P. Robin Hiesinger[1]*

- Growth cones show biased stochastic search with significant overlap
- Growth cones search is independent of target cell
- Self-organised filopodia meshwork based on neighbour interactions





“Incoming” vectors from 6 separate bundles establish neural superposition

“Outgoing” vectors from the same bundle

E. Agi et al and R. Hiesinger *Science*. 2024 Mar 8;383(6687):1084-1092
M. Langen et al and R. Hiesinger. *Cell*. 2015 Jul 2;162(1):120-33

Conclusions

- **Genetic code**: deterministic, requires mechanisms for error minimisation (**proofreading and « smooth encoding »**)

- **Transcriptional code**: **smooth encoding**, but also **combinatorial encoding** and integration relaxes constraints on 1-to-1 specificity, and increases repertoire of context-dependent regulation.

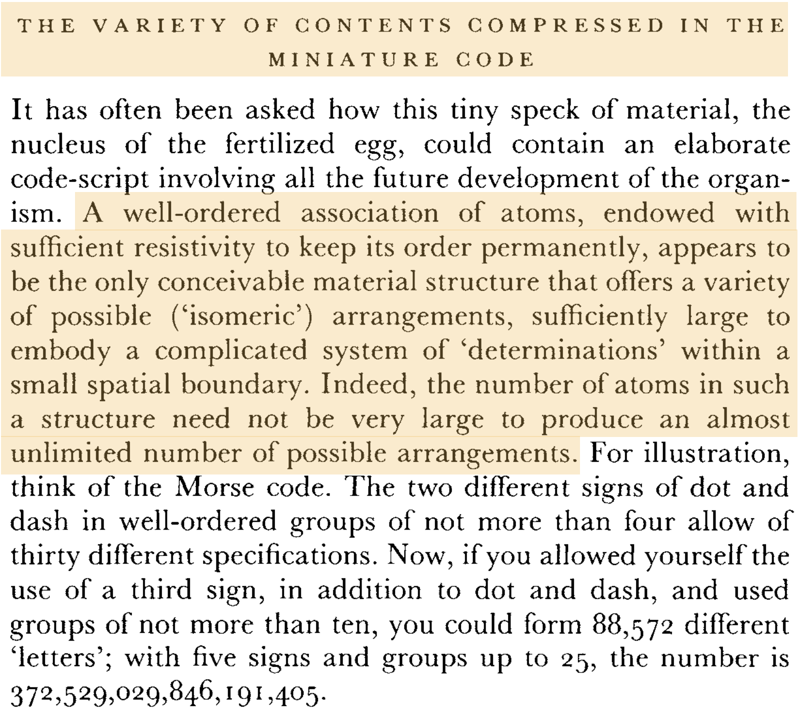
- **Signalling code:** Promiscuous binding and **combinatorial encoding** increase cellular addressing compared to 1-to-1 L/R signalling. Also allows signal computation.

- **Adhesion code**: biased stochastic processes rather than deterministic encoding. Many small contribution rather than few, selective, deterministic molecular codes.

Conclusions

Some features of biological encoding

- **Coding theory** provides a framework to understand constraints on code evolution (error load, diversity and cost). **Smooth encoding.**
- **Combinatorial encoding** increases specific « addressing » (cell identity, cell responses)
- **Deterministic use of code**: genetic code
- **Stochasticity and Algorithmic encoding**: more consistent with self-organisation.
  - Random search and stabilisation of final configuration based on energy minimisation (DAH)
  - Biased stochastic search and final stabilisation by target (neural superposition).



THE VARIETY OF CONTENTS COMPRESSED IN THE MINIATURE CODE

It has often been asked how this tiny speck of material, the nucleus of the fertilized egg, could contain an elaborate code-script involving all the future development of the organism. A well-ordered association of atoms, endowed with sufficient resistivity to keep its order permanently, appears to be the only conceivable material structure that offers a variety of possible ('isomeric') arrangements, sufficiently large to embody a complicated system of 'determinations' within a small spatial boundary. Indeed, the number of atoms in such a structure need not be very large to produce an almost unlimited number of possible arrangements. For illustration, think of the Morse code. The two different signs of dot and dash in well-ordered groups of not more than four allow of thirty different specifications. Now, if you allowed yourself the use of a third sign, in addition to dot and dash, and used groups of not more than ten, you could form 88,572 different 'letters'; with five signs and groups up to 25, the number is 372,529,029,846,191,405.