### **What is biological information?** article of  $\mathbf{C}$



#### Course 2: Biological codes Figure 12.13(A). The second example shown in Figure 12.13(B) is very familiar  $\frac{1}{2}$  cartoons in which a specific linear order of processes units under order of processes under order of processes under order orde

of what has been called the Thomas Lecuit  $T_{\rm F}$  in Figure 12.13(C) is branched assembly, such assembly, such assembly, such as the assembly, such assembly,

chaire: Dynamiques du vivant be put to put to the pieces to the piece



- 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:





Thomas LECUIT 2024-2025





Samuel Morse (1791-1872)

## • 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.





### Biological Information is mostly chemical Mean speed 31.3 *µm/s*

 $125.8\pm2.0$  and  $20.8\pm2.0$  and  $20.2\pm3.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$ 

Reading and decoding information from the environment during chemotaxis



oing up the gradient than down, chemoattractants. The binding of chemoattractants sets of a signaling cascade in the transduction module that culminates in the phosphorylation of the messenger molecule CheY. In the actuator module, the interaction of CheY-P with the fagellar motor alters the so they go up the gradient adie • Cells spend more time going up the gradient than down,



Thomas LECUIT 2024-2025 **R. Phillips, The Molecular Switch: signaling and allostery.**<br>
Thomas LECUIT 2024-2025 **References of the sequences of runs we conclude that the sequence of the sequence of r** *Princeton Univ. Press.* 2020  $\mathbf{F}$  of autocorrelations of autocorrelations of sequences of

# Reading and decoding information from the environment during chemotaxis



• Specificity: glucose, ribose, INPUT INPUT EXAMPLE A SUMPLE SERVICE ASSAULT A SERVICE ASSAULT A LOCAL SERVICE AND SUMPLE SERVICE ASSAULT A LO

### • 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









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



Thomas LECUIT 2024-2025

## 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)



Thomas LECUIT 2024-2025

**The Bell System Technical Journal**

*Vol. XXVII J Illy, 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<sup>1</sup> and Hartley<sup>2</sup> 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

<sup>-1</sup> Nyquist, H., ''Certain Factors Affecting Telegraph Speed,*'' Bell System Technical Jour-*<br>Jual, April 1924, p. 324; ''Certain Topics in Telegraph Transmission Theory,'' *A. I. E. E.*<br>Trans., v. 47, April 1928, p. 617.

<sup>2</sup> Hartley, R. V. L., ''Transmission of Information,'' *Bell System Technical Journal*, July<br>1928, p. 535.

379

# 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





R. Phillips, The Molecular Switch: signaling and allostery. Princeton Univ. Press. 2020  $\Gamma$  indiction of a signalized in the transduction module that culminates in the transductio

• Information *transfer in a noisy channel*



12

What is relevant information? ่<br><br>าล†i

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



I. Biederman, *Psychological Review* 94, 115–147, 1987<br>13 Recognition by components theory



the components were not removed, as with the circular rings of

What is relevant information?





OLJf =

 $\overline{\Pi}$  $\overline{\Pi}$  $\frac{\Box}{\Box}$ 

OLJf  $\Rightarrow$ 

I Diodorman *Devehological Deview* 04, 115, 14  $t$ . Diederman, Tsychological Review 54,  $113-14$ 

I. Biederman, *Psychological Review* 94, 115–14<sup>'</sup> contours have been deleted in regions where they can be replaced

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

#### ant biological inform tion? Therefore, we protein that, all else being equal, highly abundant proteins but have unknown or poorly ch and that table come from? Do they make sense? What do they imply about What is relevant biological information?

**values for a bacterial** *E. coli* **cell, the single-celled eukaryote :<br><b>11Ga na Whallan C.a MUhENdO GulliS**se are crude<br>ppily dividing cells of the common lab strains. Table 1: Typical parameter values for a bacterial *E. coli* cell, the single-celled eukaryote S.<br>cerevisiae (budding **vea. 41 G. 1917)**<br>characteristic values for bannily dividing cells of the common lab strains cerevisiae (budding yeast), and a mammalian HeLa call line. Motel that these are crude characteristic values for happily dividing cells of the common lab strains.



Many previous analyses focused on proteins that are expressed at



sense are they "different"?







sum<mark>e in the crowded intervolume interveneral costs are roughly independent of the crowded intervent of the c</mark>ro (specifically, HeLa: V  $\approx$  3000  $\mu$ m<sup>3</sup>; L  $\approx$  20  $\mu$ m;  $\tau$   $\approx$  1 day)



In the near function  $\mathcal{L}$  and  $\mathcal{L}$  are near function  $\mathcal{L}$  and  $\mathcal{L}$ 

yeast S. cerevisiae, and a mammalian cell adherent HeLa cell). R. Milo and R. Phillips Cell Biology by the numbers. *Garland Science*

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



# What is relevant biological information?

### Genome by the numbers

## Complexity does not scale with genome size/gene number





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



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



- *Figure 16.* Example of five stimulus objects in the experiment on the  $\bullet\,$  What are the relevant effective tuning variables? • What are the relevant effective tuning variables?
- $\bullet\,$  Ex: cell actin cortex tension and material properti • Ex: cell actin cortex tension and material properties depend on 100s of proteins
- $-$  few mechanical parameters such as stiffnes familiarity, performance in the non-recoverable conditions in the non-recoverable conditions in the non-recove<br>The non-recoverable condition was also condition was also condition was also condition was also condition was tact versions. The middle column shows the recoverable versions. The — few mechanical parameters such as stiffness and viscosity.
- $-$  few molecules with key regulatory function.  $\epsilon$  iew more rationary poor, with  $\epsilon$ contours have been deleted in regions where they can be replaced through collinearity or smooth curvature. The right column shows that  $\mathcal{L}$ — 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<sup>4</sup> 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?



### mRNA/tRNA-aminoacyl:

how to encode amino acid recruitment in protein synthesis?



# Word code: codon Message: amino acid

✐



Thomas LECUIT 2024-2025

#### Genetic information flow subsets and versions  $\mathsf{G}_\mathsf{S}$  $\mathbf{u}$ it completely in this article would be impossible. In this case, we can a ately limited myself here to presenting a broad general view of the problem,

#### ON PROTEIN SYNTHESIS great interest, has been omitted because its implications are not clear. molecular biology-those of gene action and nucleic acid synthesis. In

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 polystecharies. The main junction of process is to act as enzymes.<br>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. by polysaccharides. The main function of proteins is to act as enzymes.

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

#### and rather simple, and this fits in with this fits in with the modern view that generalizes  $\alpha$ lysis to being based upon the nucleic acids. The nucleic acids in a native protein the nucleic that is good evidence that in a native protein the nucleic that is good evidence that in a native protein the nucleic that is g

The Doctrine of the Triad.

nation of virtues has been clearly grasped.

The Central Dogma: "Once information has got into a protein it ean't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we <u>may</u> be able to have  $\frac{1}{2}$  is impossible of water. Apart from minor exceptions (such as S-S-S links form the covalent bonds of the protein is always the same of the same  $\frac{1}{2}$ 



FH. Crick. Symp Soc. Exp. Biol. 12:138-163 (1958) for the article published as Crick FHC [1958] *Symp Soc Exp Biol* XII:139-163. The



DE FRANCE Thomas LECUIT 2024-2025

#### The *ru* region consists of two adjacent genes, or 'cistrons', called cistron *A* and cistron *B.* The wild-Discovery of the Genetic Code for Proteins B) and on *E. coli Kl2* (:A) (here called K), but a phage

#### Proceedings of the NATIONAL ACADEMY OF SCIENCES re-named *FO* 0) in the Bl segment of the *B* cistron.  $$ as proflam and mutagens because the set

which has lost the function of either generators  $\mathcal{L}_{\mathcal{A}}$ 

volume 43 - Number 8 - August 15, 1957 favour of this is that mutants produced by acriding  $V$  of this is that mutants produced by acriding  $b$ 



(c) The sequence of the bases is read from a :fixed

The evidence that the genetic code is not over-

..<br>ON THE IMPOSSIBILITY OF ALL OVERLAPPING TRIPLET CODES **IN INFORMATION TRANSFER FROM NUCLEIC**  $ACID$  TO PROTEINS and mutants made with base analogues and  $ACID$  the mutants made with base analogues and  $ACID$  $\mathbf{B}$ re leaky); (2)  $\mathbf{B}$ re S. Brenner  $\emph{DN THE IMPOSSIBILITY OF ALL OVERLAPPING TRI}$ 

MEDICAL RESEARCH COUNCIL UNIT FOR THE STUDY OF THE MOLECULAR STRUCTURE OF mutants of the lysozyme of phage *T4* produced by BIOLOGICAL SYSTEMS, CAVENDISH LABORATORY, CAMBRIDGE, ENGLAND  $\,$  MEDICAL RESEARCH COUNCIL UNIT FOR THE STUDY OF THE MOLECULAR ST

Communicated by G. Gamow, June 10, 1957





LEGE<br>RANCE Thomas LECUIT 2024-2025

• 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. BIOCHEMISTRY: S. BRENNER,  $\overline{n}$  there shares  $\overline{z}$  ht with hext triplet. jondrady. Or dipides degendrated in 20 da.

- There could not be more than 256 dipeptide sequences (represented by sequence of 4 nt). Yet there are in theory 400 dipeptide sequences. **•** Therefore, peptide sequences would be constrained: refore, peptide sequences would be constrai c coald not be more than 250 dipeptide sequence  $\mathcal{L}_{\mathcal{A}}$  sufficient sequences are known to prove that it is impossible to code. here are in theory 400 dipeptide sequences.  $\hphantom{1}$ special way of degenerating the triplets. It consists in the demonstration that
- *Proof* based on data (*reductio ad absurdum*): <mark>64 triplets are</mark> insufficient to code the known aa sequences. pased on data (reductio ad absuldum). O<del>T</del> incient to code the known aa sequences.

Any triplet can be preceded (or succeeded) by only 4 different nucleotides, hence 4 different triplets. The counted from a table of neighbors.  $t_{\text{t}}$  must have preceded to succeded, by only  $\tau$  direction

Consider  $j,k,l$  aa. For every triplet for  $k$ , there are at most  $4j$  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. ider  $j$ , $\kappa$ , $\iota$  dd. For every triplet for  $\kappa$ , there are at most  $4$   $j$  iv-i



#### Discovery of the Genetic Code for Proteins  $\mathbf{C}^{\bullet}$  the bases is read from a sequence of the bases is read from a  $\mathbf{C}^{\bullet}$ sequences of bases are to be correctly read of  $\mathcal{L}$ which has lost the function of either generation of either generation of either generation of either generation of  $\mu$ covery of the genetic  $\mathbf{b}$ ode for Proteins  $\theta$  for Drotoing ae for Proteins are present

This is the phenotype shown by a complete deletion

double mutant *FO* (0 + 1), then although the

the gene and they are all non-leaky *r* mutants.

at the wrong point) in two different ways, depending  $\alpha$ 

are also found. Other mutations, known as 'leaky', show partial function; that is, they will grow on *K* 



lacking in the function of the generator of the gene. Since  $\mathbf{F}_{\text{c}}$ was published, experimental data from two sources of two sources  $\mathcal{L}$ 1)  $\alpha$  -shift is made; since

23 about half the mutants made with base analogues are leaky); (2) Street in that whereas found that whereas found that whereas found that whereas found that whereas  $\sim$ lacking in the function of the gene. Since our note was published, experimental data from two sources  $\mathcal{L}$ have been added to our previous evidence: (I) we

COLLEGE<br>{ DE FRANCE Thomas LECUIT 2024-2025

produced by nitrous acid. In an overlapping

chamed the chamed on the alternations produced in the alternations produced in the alternations produced in the second intervals of the second in the second intervals protein of the virus show that usually only one

The evidence that the genetic code is not over-

select the right triplets. If the starting point is

 $\frac{E}{\epsilon}$   $\frac{DE \text{ H} \cdot \text{H}}{1530}$  momas  $E$  toon  $E$  $\frac{1330}{20}$ triplet code, an alteration to one base will in general

 $C. O. L. E. G. E$ 

#### Discovery of the Genetic Code for Proteins  $\overline{a}$ *<u>Find</u> CODE TOTELLIS -m,* modulo *n*  **Discovery of the Genetic Code for Proteins**

#### where *n* (a positive integer) is the coding ra.tio (that verlapping triplets of bases and *m* is any integral number of bases, positive or material. A simultaneous infection of *K* by the  $\bullet$  The code is made of non overlapping triplets of bases  $\bullet$  The code is made of non overlapping triplets of bases



must be put to genetic together in the same piece of genetic same piece of genetic same piece of genetic same <br>The same piece of genetic same piece of genetic same piece of genetic same piece of genetic same piece of gene

Figure 4.9: Cartoon showing the restoration of phase in DNA sequences through R. Phillips, J. Kondev, J. Thériot & H. Garcia. *Physical Biology of the Cell (Garland Science*) 2012 in the same piece of DNA, as in the pseudo-wild



• Proflavin leads to addition of base (+); or deletion of base (—) Hender type gene both give the r phenotype (no plaque, *ie*. non functional T4 on *E*. coli K strain) • Proflavin leads to addition of base (+); or deletion of base (-) and first approximation) would be obtained in which which in which we can also be obtained in which which we obtain  $t \rightarrow t$ eds to addition of base  $(+)$ ; or deletion of base  $\ell$  $\det G$  addition of base (+); or deletion of base  $\alpha$  prienciple (no plaque, i.e. non functional  $\alpha$  is

Double Mutants

- represents

- Results:
- + with reverts to wild type Now our theory clearly predicts that all combina-
- $\frac{p}{q}$  with  $\frac{p}{q}$  mairitain r phenotype
- + ; +; + reverts to wild type





Table 3. TRIPLE MUTANTS HAVING A WILD OR PSEUDO-WILD PHENO-<br>  $TCPE$ <br>  $F(A \cup A \cup A \cup A)$ 

TYPE  $FC (0 + 40 + 38)$ <br> $FC (0 + 40 + 58)$  $(0 + 40 + 58)$  $FC (0 + 40 + 57)$  $FC (0 + 40 + 54)$ <br> $FC (0 + 40 + 54)$  $FC (0 + 40 + 55)$  $+21 + 23$  $FC(0 + 40 + 55)$ <br> $FC(1 + 21 + 23)$  $\alpha$  in our way, but incorrectly (by starting  $\alpha$ ),  $\alpha$ 

F. Crick, L. Barnett, S. Brenner and J. Watts\_Tobin. *Nature* 192: 1227-1232 (1961)  $\ldots$   $\ldots$ ,  $\ldots$  barnett,  $\ldots$  biennet and  $\ldots$  math  $\ldots$  frame

peptide chain without disturbing the function of the

#### **Enraveling the genetic code tide residues according to chain length. and further purified on Whatman 3 --Mg++ 0.09 +sRNA (deacylated) at 50 min 0.500 A"26 units 5.69 2.500 A200 units 5.36**  raveling th **(mumole base (u,umole) residues) C4\_-Phe- C14-Lys- C\_4-Pro- sRNA sRNA sRNA None 0.19 0.99 0.25**

**Oligonucleotides were eluted with H20** 

**PolyA, 16 0.22 4.35 .17** 

#### **RNA Codewords and Protein Synthesis tion was estimated by subjecting 2.5 Protein Synthesis raphy (Whatman 54 paper) both with**

**The Effect of Trinucleotides upon the Binding of sRNA** to Ribosomes **solvent A and with solvent B (40 g Proprietive disponent in 100 ml and 100 ml 7.000 ml and 100 ml and 100 ml 7.000 ml 7.0** 

**(6, 7).** 

**\_j** 

**-a' 3 O 2 -:** 

**1 3** 

**8 z 7 a: ze** ゃ **-J I CL** 

**Marshall Nirenberg and Philip Leder nucleotide were subjected to chromatog-Nirenberg and Philip Leder** 

 $\frac{1}{20}$ 

**transfer enzyme was shown to be re-** $\frac{2}{3}$   $\frac{1}{2}$   $\frac{1}{2}$ 

**four chromatographic systems described**   $\frac{1}{2}$  $\begin{bmatrix} \frac{1}{2} & 6 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$  **Contains greater in a model of the state of the s be detected. Several preparations of**   $\begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 & \mathbf{e}_5 & \mathbf{e}_6 & \mathbf{e}_7 & \mathbf{e}_7 & \mathbf{e}_8 & \mathbf{e}_7 & \mathbf{e}_$ **preparations, no contaminants were de-** $\frac{1}{2}$  **tech**<br> $\frac{37}{2}$  **1**  $\frac{37}{2}$ 

**2.7 (0.05M ammonium formate, 80** 

**2.500 A260 units 2.08** 

 $T = \frac{1}{2}$   $\frac{1}{2}$   $\frac{1}{2}$ **length of many code-**  $\bullet$  or  $\Box$  $\mathbb{E} \left\{ \left. \left\| \left( \mathbf{r} \right) \right\| \right\| \leq \mathbf{r} \right\}$ **of chemically defined oligonucleotides**   $\frac{1}{2}$   $\frac{3}{2}$   $\frac{1}{2}$   $\frac{1}{2}$  **to rigosomes**  $\mathcal{F}$ **method of detecting the detection and**  $\mathbf{r}$  $\left[ \frac{37^{\circ}}{40^{3}} - \frac{37^{\circ}}{40^{3}} \right] - \text{POLY}$ 

To determine the minimum chain length of mRNA required for codeword recognition and to test the ability **polynucleotides containing randomly or-of chemically defined oligonucleotides to induce C<sup>11</sup>-aminoacyl-sRNA binding** to ribosomes, we have devised a rapid method of detecting this interaction and have found that trinucleotides are ac**structure. Since oligonucleotides of tive as templates. known base sequence are readily pre-To determine the minimum chain** 

**in many ways, to use defined oligonu-**Effect of polyU upon the rate of C<sup>14</sup>-Phe-transfer tRNA binding to ribosomes. To ober **direct which is a simple of our contract of our contract of oligonucleotides. The simulation of or**  $\mathcal{L}$ **should provide a general method for**   $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ **the proportions indicated: No. 591,** 

**triplets of known sequence. The system**  ic dimacreodiaco, popo<sub>l</sub> but not dinucleotides, direct the binding to ribosomes **mRNA which occur during the process**  f phenylalanine-, lysine-, L nAnAnA, and nCnCn **tain oligonucleotides with different polyC (12) were digested as follows: but not amacrootides, an editine smallig to his sort**<br>of phenylalanine-, lysine-, and proline-tRNA. **phosphate; 100 mg of polynucleotide were incubated at 37?C for 18 hours in**  The trinucleatides, pulpl **polyC (12) were digested as follows: a 28-ml reaction mixture containing**   $11nA<sub>0</sub>A<sub>0</sub>$ **Base composition and position of digesting 2.0 A260 units of the CDM** of phenylalanine-, lysine-, and proline-tRNA.<br>. The trinucleotides,  $pUpUpU$ ,  $pApApA$ , and  $pCpCpC$ ,



**ported by Cannon, Krug, and Gilbert (4). However, the addition of polyU** 

**Thomas LECUIT 2024-2025** 



Fidelity in protein synthesis. Many difer- $\epsilon$  $\sum_{\text{max}}^{24^{\circ}}$ 

**MINUTES** 

**portions of buffer at 0? to 3?C. Ribo-**

of those species is the correct one.

**After each fraction of each interval state where**  $\mathbf{R}$  **sandwich where**  $\mathbf{R}$  **sandwich where**  $\mathbf{R}$  **satisfying the same of state**  $\mathbf{R}$  **satisfying the same of state**  $\mathbf{R}$  **satisfying the same of state**  $\mathbf{R}$  **s CN C** (1927-2010)



M Nirephers and  $\overline{P}$  Leder Science 145: 1300 1407 (1064) M. Nirenberg and P. Leder. *Science* 145: 1399-1407 (1964) **units of T2 ribonuclease (16) in 0.1M** 

# 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



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



« 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 effect overy the thing compressing the jetting (i.e. the case of bacterial replication) in 40 minutes. In this incomponent delivery process, the truck would deliver a wrong package only once every 3 years! »  $\alpha$  is comes as no surprise that a big effort has been made to build specific specific specific specific specific

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



#### in 104 Dealing with errors but similar amino acid is inserted has been estimated at <sup>1</sup> maintained in the, several recognition steps between amino $s$  $\sigma$  that necessary for the operation of this mechanism.

### **External General General Property of Algebra**  $\bullet$  Kinetic Proofreading stand how small error rates are achieved.

Substrate *C* (resp. *D*) by recognition site *c* (resp.*d)*: Michaelis kinetics. The reactions



John Hopfield (1933)

Kinetic Proofreading: A New Mechanism for Reducing Errors in **Biosynthetic Processes Requiring High Specificity** (protein synthesis/DNA replication/amino-acid recognition)<br>incorporated have essentially undistinguishable energies.

J. J. HOPFIELD J. J. HOPFIELD<br>Department of Physics, Princeton University, Princeton, New Jersey 08540; and Bell Laboratories, Murray Hill, New Jersey 07974<br>Pepartment of Physics, Princeton University, Princeton, New Jersey 08540; and Be

acid monomer and the product protein. Indeed, one of the

#### Kinetic amplification of enzyme discrimination. • • Fauil ferences in intermediates or kinetic barriers by a process

 $\rho_{\rm AGQ}$  and  $\rho_{\rm AGQ}$  and  $\rho_{\rm AGQ}$  . The Salk Institute for Biological Studies P.O. Box 1809,<br>San Diego California 99119 acid recognition, and DNA replication, all exhibit the *San Diego, Ca!i[ornia 92112.*   $f(12-12-1974)$ .  $(12-12-1974).$ 

J.J. Hopfield (1974) PNAS (10): 4135-9 J. Ninio (1975). Biochimie. 57 (5): 587-95. maintained in the, several recognition steps between amino $t - t$ 

 $k'c$  $C + c \rightleftarrows$  Cc  $\rightarrow$  correct product  $K_C = k'c/k_C$ kc  $k'$ <sub>D</sub> W  $D + c \rightleftarrows Dc \rightarrow \text{error product} \quad K_D = k'_{D}/k_D$ kD

iscrimination. • Equilibrium discrimination is not sufficient to same precision as the conventional first identification, the  $\frac{1}{\text{c}}$ account for measured error fates.

Error rate =  $p_D/p_C = K_C/K_D = e^{-\beta \Delta E_{CD}}$ Energy difference scales with energy associated with formation of  $5-9$  1 H-bond:  $\Delta \varepsilon \approx 2k_BT$ and  $D$  by a recognition site caforing  $\mathcal{L}$ Error rate  $\sim$ e $\sim$ 2 $\sim$  0.13 For measured error rate  $\sim$ 10<sup>-4</sup>  $\frac{1}{2}$ indiscriminant incorporation is reasonable when the covalent values of  $\frac{1}{2}$ Energy difference scales with  $s \sim \text{softmax}$  in  $s = \text{length}$ 

> $\Delta \varepsilon \approx$  - $k_BT\ln 10^{\text{-}4}$   $\sim$   $10k_BT$  $\Delta c \sim \gamma_B T$  into  $\gamma$  ionsistent



Thomas LECUIT 2024-2025

#### complete the Dealing with errors source might in a typical example use  $\mathsf D$  ealing with errors  $\mathsf D$ maintained in the, several recognition steps between amino- $\mathcal{T}_{\text{max}}$  $\mathbf h$  and with errors Define the error fraction f as the rate of incorrect product product product  $\Box$  $T_{\rm F}$  for the reactions  $T_{\rm F}$  for the error fractions  $T_{\rm F}$  for the error fraction  $T_{\rm F}$

#### **Extending Correct Correct Proofreading** • Kinetic Proofreading **Construction Construction Construction**  $T_{\rm eff}$  for the error fraces in  $5.8$  for the error fraces in  $\sim$  $T_{\rm tot}$  scheme for discrimination between substrates  $C_{\rm tot}$  and  $C_{\rm tot}$  $\mathbf{F}$  reference purposes, no special suppositions about "on"  $\mathbf{F}$  suppositions about "on"  $\mathbf{F}$ when substrate  $C$  and  $D$  are in equal concentrations. For the re- $\mathbf{b}$ constraint allows far better error fractions. Suppose  $\mathcal{L}$

 $f(x + k)$ <br>with e anocificity (i.e. discrimination)  $t_1$  is in which is in which  $\frac{1}{2}$  is a important, it is desired in  $\frac{1}{2}$  is desired in  $\frac{1}{2}$ • on rates are the same:  $k'c = k'D$ So off rates carry the specificity (ie. discrimination).

fundamental general problems of biosynthesis is to under-

• Minimum error attainable :  $\bullet$  ivinimum error d

for a formation divided by the rate of correct product for  $\sigma$ 

 $k_c/k_D = K_D/K_c \equiv f_0 = \exp -(\Delta G_{CD}/RT)$  $\kappa_{C}/\kappa_{D} = \Lambda_{D}/\Lambda_{C} = 0$  is exp  $-\left(\Delta v_{CD}/\Lambda_{I}\right)$ 

 $\overline{\phantom{a}}$  define the error fraction f as the rate of incorrect product prod

- $\frac{1}{100}$  the stage kinetic model which is the two-stage two-stage kinetic model which it is the stage of the stage  $\frac{1}{100}$  $t$  the same kind of discrimination. The reaction pathway for  $t$  and  $t$  and  $t$ energy consuming anven n  $\bullet$  Introduction of high energy intermediate Cc\* produced by energy consuming driven reaction (ie. GTP hydrolysis) (2).<br>
Set discovistes we we should then  $\mathsf{D}_{\mathsf{S}}\mathsf{t}$ 
	- ore slowly than Dc<sup>3</sup> greater than or experience that the AGCD/RT is contained to experience and the AGCD/RT is contained to the AGC •  $\mathsf{Cc^*}$  dissociates more slowly than  $\mathsf{Dc^*}.$

ates more slowly than DC .<br>cinetic model iterates the same discrimination.<br>at *m*' is substrate independent and that ۔<br>من<del>ا</del>موہ  $1$  and equilibrium ) the error rate is:<br>equilibrium ) the error rate is: • Cc alssociates more slowly than *Dc* .<br>• Two stage kinetic model iterates the same discrimination. Assuming that *m'* is substrate independent and that  $\frac{1}{2}$  is the  $\frac{1}{2}$  kJ is the  $\frac{1}{2}$  calculated to  $\frac{1}{2}$  and  $\frac{1}{2}$  is the  $\frac{1}{2}$  $m'$  <  $k_C$  and  $W$  <  $l_C$  , (ie. reactions 1 and 2-3 are at near  $m < k_C$  and  $w < l_C$ , (ie. reactions 1 and 2-3 are at near<br>equilibrium ) the error rate is:<br> $f=f_{step1} x f_{step2} = k_C/k_D x l_C/l_D$ 

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

the same « reading » mechanism is used for dissociation from  $Cc^*$  and  $Dc^*$  as from  $\int$  *Cc* and *Dc then:*  $f = f_o$  *<sup>2</sup>*  $\frac{1}{2}$  for a ratio for the entrance to the second in-Because the only simple discrimination mechanism is a LAGCD if the same « reading » mechanism is used for di  $\mathbf{C}$  can be based on the kinetic term in  $\mathbf{C}$ If the same « reading » mechanism is used for dissociation from *Cc\** and *Dc\** as from same «<br>d De th near equilibrium ratio between  $Cc^*$  and  $Dc^*$  as from  $\mathcal{O}(\mathcal{O}^2)$  substrate are present. Thus  $\mathcal{O}(\mathcal{O}^2)$ 

c and Dc then,  $J - Jo$ <sup>-</sup> By adding *n* steps, discrimination is increased with error rate  $f = f_o^{n+1}$  $r_{\text{r}}$  behaves  $r_{\text{r}}$  behaves in a fashion analogous to  $r_{\text{r}}$ 

 $u \in \mathbb{R}^n$  in a situation where the source population where  $\mathbb{R}^n$ **J.J. Hopfield** (1974) *PNAS* (10): 4135–9

must be met. First, the wrong substrate arriving at DC\*





 $k'c$  intermediate, so that leads that leads that leads the  $lD'$  $T_c \leftrightarrow C_c \rightarrow$  correct product  $\Lambda_c = \frac{\kappa}{c} \frac{\kappa}{c}$  $k'$ phosphate hydrolysis. The back reaction m can be back re  $m\tau\rightarrow Dc\rightarrow e$ rror product $\overline{K}_D=k'{}_D/k_D$  $\kappa$ *b*  $k'c$  IV  $C + c \rightleftarrows Cc \rightarrow \text{correct product}$   $K_C = k'c/k_C$  $k'$ <sub>D</sub> W  $D + c \rightleftarrows Dc \rightarrow \text{error product} \quad K_D = k'_{D}/k_D$  $\mathbf{c} \rightarrow \text{correct product} \quad K_{\textbf{C}} = k'c/k_{\textbf{C}}$  $\mathcal{W}$  and multiplier  $W$ . With this constraint,  $W$  $a \rightarrow Dc \rightarrow \text{error product}$  $\mathbf{k}_D$  in this system requires energy for  $\mathbf{k}_D$ reasons sketched in the following section. Let the intermedi $k^i c$  intermediate, so that leads that leads the 1D's negligible. ).  $\frac{C}{c}$  of  $\frac{C}{c}$  comes from the driven reaction  $\frac{C}{c}$  $c \overset{k'D}{\rightleftharpoons} bc \overset{W}{\rightarrow} \text{error product}$   $K_D = k' \frac{E}{D} k_D$ made negligible by  $k_{\text{D}}$ 

and pyrophosphate as the product f3.



#### rates, the single intermediate in a  $\Box$ have this property, and acts as if  $1c$   $\geq$ is sufficiently specific that it was used (14, 15) in deciphering Dealing with errors  $T$  enzymatic binding process involves the process involves the process involves the prior formation  $\mathcal{D}$

#### In as soon as cC is formed, it is formed product at a rate W. In  $\mathcal{L}$ , the formation of C. In  $\mathcal{L}$ of a terminal complex of the transformation factor  $\mathcal{A}_\mathbf{R}$ • Kinetic Proofreading and translation and the Tu factor is released. The total released. Thus, the total reaction scheme  $\sigma$

 $\bullet$  The aminoacyl-tRNA is activated via GTP by chain (by EF-Tu in prokaryotes or eEF1A in eukaryotes). • The aminoacyl-tRNA is activated via GTP hydrolysis before incorporation to the aa Biology: All the  $\sim$  Biology



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



problem in collapsing  $\mathcal{A}$  on  $\mathcal{A}$  is the representation of  $\mathcal{A}$  onto  $\mathcal{A}$ 





12.2 A Panoply of Demonic Behaviors in the Living World 375

**relation for the R. Phillips The Molecular Switch is not relation in the R. Phillips The Molecular Switch is no** statistical mechanics of competitive binding, where for simplicity we imagine J.J. Hopfield (1974) *PNAS* (10): 4135–9R. Phillips The Molecular Switch. Princeton Univ. Press

#### e Kingtic Proofreading and translation have the this property, and acts as if 1c  $\sim$  $\bullet\,$  Kinetic Proofreading and translation  $T_{\rm eff}$  enzymatic binding process involves the process involves the process involves the process involves the prior formation  $\sim$  $125.8\pm2.0$  and  $20.8\pm2.0$  and  $20.8\pm2.0$  and  $20.2\pm2.0$   $\pm2.0$   $\pm2.0$   $\pm2.0$

occur. Nonenzymatic'binding at <sup>a</sup> <sup>20</sup> mM Mg concentration,

 $\tau$ In  $\tau$  is to make the make  $\bullet~$  There is a large excess (few 10 fold) of wrong tRNA-aa competing with a produced the generation, but in the generation given correct tRNA-aa for binding to an anticodon.  $\bullet~$  There is a large excess (few 10 fold) of wrong tRNA-aa competing with a  $\qquad \qquad \blacksquare$ 





- From the in the range of  $10^{-4}$  -10<sup>-6</sup> /codon should inevitably produce  $\mathcal{L}_{\mathcal{A}}$  , but with an effective time-dependent W(t) (the measure-dependent W(t) (the measure-dependent W(t) (the measureproteins 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<sup>10</sup> proteins/cell, so there should be many tion of Cc in an otherwise normal Michaelis discrimination. proteins with wrong incorporated aa. **But this is not the case.**  $\bullet~$  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 sized protein. There are 10<sup>10</sup> proteins/cell, so there should be many  $t$  system adopts a distribution of all products and reactants, as shown in  $\alpha$ Figure 12.13(A). The second example shown in Figure 12.13(B) is very familiar
- $\bullet~$  How does the cell cope with this given the impact on function?  $\bullet~$  How does the cell cope with this given the impact on function?
- $\blacksquare$  How to minimize this error load?  $\bullet\;$  How to minimise this error-load?

mediate, Cc. Since both intermediates in [9]. can break up



## • 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.



32

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

- 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)
- T. Tlusty, *Physics of Life Reviews* 7 (2010) 362–376 • Generic model of genetic code evolution:
- 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)

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

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



Thomas LECUIT 2024-2025

• Coding transition is governed by properties of this information channel

 $fithess = -error$ *-load* +  $w<sub>D</sub>$  *×diversity* −  $w<sub>C</sub>$  *×cost* 



- *•* 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.

Thomas LECUIT 2024-2025

COLLÈGE

 $p_{ia}$ = probability that codon  $i$  matches amino acid  $\alpha$ The  $N_c \times N_A$  probabilities form a **code-matrix.** Initially the association is random and all  $p_{ia} = 1/N_A$  (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<sub>C</sub>* vectors in an *N<sub>A</sub>* dimensional space. <sub>"</sub>

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

*pi<sup>α</sup>* + *pi<sup>β</sup>* + *pi<sup>γ</sup>* = 1 Blue: codon encodes a single aa Green: codon encodes 2 aa Red: codon encodes 3 aa (no information)

```
T. Tlusty, Physics of Life Reviews 7 (2010) 362–376
Fig. 2. The genetic code as a noisy information channel. (A) The genetic code is a mapping that relates the space of the NC = 64 codons i, j, k,...
```
• Coding transition is governed mainly by properties of this information channel

 $fithess = -error$ -*load* +  $w<sub>D</sub>$   $\times$ *diversity* −  $w<sub>C</sub>$   $\times$ *cost* 

- The optimal code is found by maximizing the fitness with respect to the code matrix *piα.*
- 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  $w_C$



# 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

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



"Gene regulation for higher cells:

TF/DNA sequences:

how to encode gene expression of ~few 10<sup>4</sup> genes from set of transcription factors (TFs)





by the binding mechanism of that super-family. We also that super-family. We also that super-family. We also

- 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 superfamilies with larger set of TFs.



Thomas LECUIT 2024-2025

### Coding limits on the number of transcription factors

(Ga with a  $\sim$  1.8+/-- 0.17). The  $\sim$  1.8+/-- 0.17

Shalev Itzkovitz<sup>1,2</sup>, Tsvi Tlusty<sup>2</sup> and Uri Alon<sup>\*1,2</sup>  $T_{\text{max}}$  of degrees of  $T_{\text{max}}$  and  $T_{\text{min}}$ 

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





The kingdom in which each super-family is observed is abbreviated as A — Archea, B — Bacteria, E — Eukaryotes. Estimates for the number of which ease was apported by the moth of the moth of the moth of the moth of the moth O – number of possible orientations, H – homo-dimers (1) or hetero-dimers (2). The number of sequences is 4P\*H\*O\*S/2.

estimated number of possible sequences

of transcription factors from each super-family in diverse organisms. The contractors of the contractors of

estimated as 46/2 = 2048, more than the number of

# 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)







 $\mathsf{a}$ 

#### $\mathcal{C}_{\text{max}}$   $\mathcal{C}_{\text{train}}$ ,  $\mathcal{D}_{\text{train}}$ Case Study 2: Transcriptional regulatory code colored spheres. Colors correspond to the biological function of each TF. a) A sphere-packing code – code-words are covered

Protein/DNA recognition: a Sphere-packing code. rotein/DNA recognition: **a Sphere-packing co** have similar biological function, represent

by non-overlapping spheres. The TFs do not share binding sequences. A smooth code-words are covered by over-words are covered by over-

• Balancing Diversity and Specificity

The coding problem is akin to a sphere packing code in sequence space. load caused by extension in the cause of  $\beta$  $\mathbf{S}$  such codes are called  $\mathbf{S}$ .

The size of the sphere of a given TF reflects the extent of binding to *adjacent* sequences  $\mathcal{L}_{\text{eff}}$  this this this this this this takes the this takes the theory takes the theory takes takes the theory tak of the sphere of a given TF reflects the ext $\epsilon$ tantly, the theory predicts that neighbouring "spheres",

Increasing diversity will lead to overlapping spheres. that is TFs with similar binding sequences, would tend to g diversity will lead to overlapping sphere

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

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



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



Thomas LECUIT 2024-2025

Diversity: +/-**Machinary similar pictures with similar pictures of the similar picture of the similar picture of to have similar picture of the similar picture of the similar picture of the similar picture of the similar p** of **Specificity:** ++ *biological functions* that transcription factors with similar recognition sequences should tend to have similar biological effects. The reasons with similar sequences can sequence similar sequences can sequence similar sequences can sequences sometimes bind to each others sequences. If the bind to each other sequences. If the bind the bind the bind th e and and and and and and B and <br>The reduction of the reduc  $\mathbf s$  in fitness caused by such extensions would be smaller. Hence,  $\mathbf s$ there may be a selection pressure to allocate similar similar similar similar similar similar similar similar sequences to biological ly similar factors. To test the prediction that TFS with similar sequences with similar sequences with similar sequences.  $s_{\text{max}}$  function, we examine the similar function, we examine  $\sigma$ **E. colicia e. colicia e. colicia e. colicia e. compared the interest and compared the interest and compared the**<br>The contract and compared the interest and compared the contract of the contract of the contract of the cont

similarity by means of several distance metrics. organisms there exists a significant sequence similar sequence simil  $\frac{1}{2}$  some TF pairs  $\frac{1}{2}$  ,  $\frac{1$ 





S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239  $\beta$ . RENOVICE, 1. THIS can share binding sequences  $\beta$  to be indicted to be indicted to be indicted to be indiced to be i

# Case Study 2: Transcriptional regulatory code

• Hypothesis:

 $-$  Optimal coding theory predicts that overlapping spheres with  $\frac{1}{2}$ smooth coding: namely TFs with partially and the space sequence sequence should be space shown in the space space should be space shown in the space space space space shown in the space space space space space space space have similar/overlapping functions (regulate MSN)<br>minimise impact on fitness). 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, pvalue of 5.1\*10-5)



Transcription factors with overlapping binding sequences in S. cerevisae. their corresponding sets of binding sequences have significant overlap according to the present measure. Bold edges connect

S. Itzkovitz, T. Tulsty and U. Alon. *BMC Genomics* 2006, 7:239 TFs which also have biological similarity according to the functional annotation and transcription network (gene co-regulation) S. Itzkovitz, T. Tulsty and U. Alon. *BMC* C functional similarity in yeast, we used an experimentally



other words, the TF  $\sigma$  spheres  $\sigma$  spheres  $\sigma$  spheres  $\sigma$ 

### TF/DNA sequences:

how to encode gene expression of ~few 10<sup>4</sup> genes from set of transcription factors (TFs)

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

anterior–posterior (A–P) patterning of Yet, there many other layers of regulation of transcription:

- controlled by spatially varying aftacte and intagr eneew and magn • Combinatorial effects and integration on regulatory sequences (see course #3).
- $\frac{1}{\sqrt{2}}$ re dearees o  $\bullet$  Introduces more degrees of freedom and alleviates constraints from sequence overlap 100 r<br>C





Case Study 2: Transcriptional regulatory code



43

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

**COLLÈGE**<br>DE FRANCE

*ceh-28 ceh-93 ceh-8 mls-2 ceh-10 ceh-7 ceh-2 unc-39 ttx-3*

effect of loss of loss of loss of loss of specific

### *Spatial* combinatorial encoding of cell identity: vertebrate neural tube April 1988 and the April 1989 and

- 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

BMP







Malin and Desplan C. *PNAS* 2021 Vol. 118 No. 28 e2101823118 (T. Jessel and J. Briscoe labs)  $p_{A}$  and  $p_{A}$  are first protein in the expression of basic helixes the expression of basic helix (bHL) THS such as in  $p_{A}$  is such as in the expression of basic helix (bHL) TES such as in the expression of  $\mu$ 

### *Temporal* combinatorial encoding of cell identity: *Drosophila* nervous system **Type II NBs Type I NBs** nnral r

- **Thoracic NBs** • In the Ventral Nerve Chain (VNC) distinct neurons **IDED** are produced in each segment of the embryo
- **Abdominal NBs** • 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  $U_{\mathbf{u}}$

# Case Study 2: Transcriptional regulatory code

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

- 60.000 neurons in the *Drosophila v*isual system (⅔ of the brain).
- Visual system requires the specification  $\begin{CD} \mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{R}(\mathsf{H}_1(\mathsf{H}_2))=\mathsf{$ of distinct neurons.
- 100 Medulla neurons are specified from  $\begin{array}{ccc} \text{Mil} & \text{Mil} \\ \text{Mil} & \text{Mil} \end{array}$ the Outer Proliferation Center (OPC), in a  $\begin{array}{ccc} \text{Median} & \text{if } \mathcal{P} & \text{if } \mathcal{P} \end{array}$ wave.
- Each progenitor generates a clonal  $\overline{ }$ <sup>Lobula</sup>  $\overline{ }$ <sup>Lobula</sup>  $\overline{ }$ <sup>15</sup> / descent of medulla neurons organised in a column.



lobula. Distal medulla (e.g., Dm4; green) neurons are multicolumnar and  $P$  arbors according medium  $P$  and lobula  $P$ 

Malin and Desplan C. PNAS 2021 Vol. 118 No. 28 e2101823118C.  $\frac{1}{2}$  connect the lobular processes broad field  $\frac{1}{2}$  motion. Let us be a final field motion. Let us a final field motion. Let Malin and Desplan C. *PNAS* 2021 Vol. 118 No. 28 e2101823118C. neuro en to the lateral (L) side and the lateral (L) side and the lateral (L) side and to the lateral (L) side and to the lateral (L) side and the lateral (L) side and the lateral (L) side and (L) side and (L) side and (L)

Lamina

and is compared with similar mechanisms used in vertebrates.





neurons expressing both ap-LacZ (ap-Z; green), an enhancer trap that perfectly

# case study z

mimics Ap expression, and Hth (red). b, Bsh (blue), but not Hth (red), is lost in

 $\sf Case \ Study\ 2: Tran$  and  $\mathbb{R}$ *Temporal* combinatorial encoding of cell identity: *Drosophila* nervous system f ghang and ghange of the state of  $\frac{1}{2}$ 

*Orto Green Address of the Company's Comp* 

*OrtC1-gal4>GFP*

DEVELOPMENTAL BIOLOGY

- $\bullet$  A temporal cascade to TFs encode distinct cell types.  $\bullet$  A temporal cascage to TFs encode
- This is amplified by spatial cues in the outer proliferation center (OPC). outer promeration center (OTC)<br>Press Dac. The expanding documents Data
	- Notch signalling amplifies cell diversification (binary choice).

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



Pax7

superficial layers generated by Slp<sup>1</sup> and D<sup>1</sup> neuroblasts (Supplemen-



progeny of D<sup>1</sup> neuroblasts.





# Case Study 3: Signal encoding

Ligand/Receptor (signalling code): article open access to the control of the control o<br>Article of the control of the contr

how to encode specific signalling output behaviours from different ligands?





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.
- $\bullet$  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 *Promiscuous* L/R binding (many to many) Specific encoding  $\Big|\qquad\Big|\qquad$  Combinatorial encoding





Thomas LECUIT 2024-2025 >>Power and Limits of these different strategies

# Case Study 3: Signal encoding

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



Yaron E. Antebi,1 James M. Linton,1 Heidi Klumpe,1<sub>'</sub>2 Bogdan Bintu,1 Mengsha Gong,1 Christina Su,1 Reed McCardell,1<br>and Michael B. Elowitz<sup>1,3,4,</sup>\*

- Combinatorial sensing of BMP ligands.
- $\bullet$  Reports BMP signalling with fluorescent reporter using BMP responsive element  $\mathcal{L}(\mathcal{B})$ , and  $\mathcal{B}(\mathcal{B})$ , and  $\mathcal{B}(\mathcal{B})$ , and 3 type II receptors in  $\sim$ **A**





E. Antebi et al. and M.B. Elowitz. Y.E. Antebi et al. and M.B. Elowitz. Cell 170, 1184-1196 (2017)  $\gamma$ . E. Antebi et a



6

• The combinatorial logic of BMP ligands natorial logic of BM Relative activity 4

GDF6

• 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)

BMP4 only



# Case Study 3: Signal encoding

**A**

- 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)

0.2 imbalance (subtraction), balance (multiplication) • 4 core integration modes or classes of computations emerge: additive (addition), and ratiometric (division) signalling.

Thomas LECUIT 2024-2025

COLLÈGE **DE FRANCE** 



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

#### Case Study 3: Signal encoding restricted to a biologically relevant range (STAR Methods). How- $C_{\text{t,old}}$ ,  $\Omega$ ,  $C_{\text{t,real}}$  as sections  $\sigma$  othog of organizations requires  $\sigma$

- The 4 computational modes arise through the interplay between <mark>different binding</mark> affinities (allowing competition for L/R binding) and existence of different complex **Finding** activities. biological parameters could have been selected by evolution  $\bullet\,$  The 4 computational modes arise th ing migriplay between amerent binanty duuviliga. able computational diversity.
- Additive: the 2 ligands have  $\sim$  equivalent activities  $(\epsilon_{i1k} \sim \epsilon_{i2k})$  $\alpha$ cuvities (c<sub>ilk</sub> ~ c<sub>i</sub> $2k$ )
- Ratiometric: signaling complexes from one ligand have higher activities than from the other  $(\epsilon_{i1k} \ll \epsilon_{i2k})$ ing one ligand have higher activities than those containing the
- Imbalance: each receptor preferentially binds to a state of the imbalance tially binds to a distinct ligand with which it forms a less active distinct ligand with which it forms a less active complex  $\blacksquare$
- ligand with which it forms a more active complex • Balance: each receptor preferentially binds to the

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



Thomas LECUIT 2024-2025



- **•** If only 1 ligand it signals from both receptors
- **<sup>P</sup> <sup>P</sup> <sup>P</sup> <sup>P</sup>** Representative parameter regimes producing each of the four archetypes are indicated schematically. Upper and middle arrow thicknesses indicate the affinities • If 2 ligands, sorting among different receptors and lower total signalling (A) When two ligands are equivalent (similar arrow thickness  $\mathcal{L}$
- 53  $\mathcal{D}\mathcal{A}$ rise through the Internet Interactional Arise through the Interaction Affinities and Complex Activity a  $\mathcal{B}(\mathcal{B})$  when different ligands generate different levels of activity in complex with the same receptors (thin versus thin versus t

## $\frac{1}{2}$ Case Study 3: Signal encoding Relative Ligand Strength

- 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)

*Bmpr2 Acvr1 Acvr2a Acvr2b Acvrl1*



- 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.



S



### CAM/CAM (adhesion code):

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





Thomas LECUIT 2024-2025

- 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.* 









Johannes Holtfreter 1901- 1992



**NCE** Thomas LECUIT 2024-2025

### simple as to preclude the possibility of the possib Case Study 4: Cell-Cell adhesion code

#### provide the variances of the variances of the variances of "international terms of "international terms of "in  $\bullet$  Encoung ussue organisation via cell-cell adhesion energy  $\bullet$  Encoding tissue organisation via cell-cell adhesion energy

### Reconstruction of Tiss by Dissociated C

 $\alpha$  coincidents and the continuous out of embryonic cells may have a common explanation. Some morphogenetic tissue movements and the sc

3579 Malcolm S. Steinberg 704 M.S. STEINBERG AND S.F. GILBERT 2 August 1963, Volume 141, Number 3579 mon explanation.<br>alcolm S. Steinberg **SCIENCE** 



Fig. 2. Hans Holtfreter pronouncing his disagreement with the experimental evidence of Mal Steinberg (shown here<br>laughing) for the thermodynamic model of cell sorting. This photograph was taken at the embryology course at in the development of our biological control of our biological control of our biological control of our biological control of the co Biology Laboratory at Woods Hole, 1971. elicit and orient these tissue movements entiation and rediffer and that cells alterei d their cytological only the associations but also the ana ughing) for the thermodynamic model of cell sorting. This photographs elicity and oriented the models from the times. Fig. 2. Hans Holtfreter pronouncing his disagreement with the experimental evidence of Mal Steinberg (shown here<br>laughing) for the thermodynamic model of cell sorting. This photograph was taken at the embryology course at inberg their normal morphogenetic functions. 2. Hans Holtfreter pronouncing his disagreement with<br>g) for the thermodynamic model of cell sorting. This phot

erg & Gilbert *J. Exp. Zool.* 2004, about Townes & Holtfreter *J. Exp. Zool.* 1955 **WIII—-can**  $\frac{1}{2}$  $\sum_{i=1}^{n}$ same stage in vivo. In time,however,  $J. Exp.$ Steinberg & Gilbert J. Exp. Zool. 2004, about Townes & Holtfreter J. Exp. Zool. 1955



isms showld have led us to assign correlation of the state  $\mathcal{L}$ NCE Thomas LECUIT 2024-2025 Holtfreter J. 1944a. Experimental studies on the development of the pronephros. Rev Canad Biol 3:220–250. Holtfreter J. 1991. Reminiscences of the life and work of

**ues** While the *adaptedness* brought about ells through evolution appears complex, the  $R<sub>c</sub>$  anatomy of a body part may be exis quantity: more versus less. There is, analysis, their most impressive feature  $\sum_{T\text{g (shown here}}$  may be the simplicity of the terms in *adaptiveness* which makes evolution possible is born of simplicity. The entire genetic code (and more) is exinberg the state of the pressible with an alphabet containing only four elements. It would appear that a not inconsiderable amount of  $\frac{d}{dx}$  the "information" required to produce, through morphogenetic movement, the pressed in a code whose sole element I think, reason to expect that as more **could realms of biological specificity yield to**  $\mathbf{w}$  which specificity—information, if you  $\text{will—can be expressed } (34).$ ation. these fragments showed marked marked properties. the generic is only four a gastrum and  $\alpha$ same stage in time, however, however, however, however, however, however, however, however,  $\mu$  $\frac{1}{2}$  intrinsic proper- one and  $\frac{1}{2}$  and  $\frac$ Dm appeared very bythe penetration of the mesoderm be- $\mathbb{R}$  cooperations  $\mathbb{R}$  $\alpha$  sponge through isolated equal to the sponge through isolated equal to  $\alpha$ hat aggregates ober tissues to the time the permanent union of the permanent union of the permanent union of the  $\alpha$  $t<sub>trans</sub>$  sponges (1).  $t<sub>trans</sub>$  when the tissues were present in  $t<sub>trans</sub>$  were pr -d to may be the endomline position, with  $\alpha$ entiative opperation of the culture version of  $\mathcal{C}$ a<br>ic<br>ol<br>no ns<br>,e<br>e<br>e Downloaded from the from the controller in li<br>al<br>is<br>l-

DNA, at <sup>a</sup> later stage, too simple to

Steinberg MS. Science. 141:401-408. 1963  $\frac{1}{\sqrt{2}}$  $1963$ 

787 (1955).<br>1950 - Paul Barnett, politik fizikar (h. 1955).<br>1950 - Paul Barnett, politik fizikar (h. 1955).

## Differential adhesion hypothesis accounts for cell sorting in vitro



 $100 \mu m$ 



Thomas LECUIT 2024-2025

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

### 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)



Thomas LECUIT 2024-2025

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







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

- 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)

#### $\emph{CHEMOAFFINITY IN THE ORDERLY GROWTH OF NERVE}$  $FINI1$  IN THE ORDERLY GROWTH OF NERVETIBER PATTERNS AND CONNECTIONS\*  $\mathcal{B}$  and commetitors

consider them further here.

BY R. W. SPERRY

DIVISION OF BIOLOGY, CALIFORNIA INSTITUTE OF TECHNOLOGY

R.W. Sperry. *PNAS* 50(4): 703–710 (1963)  $4.50(4)$  702.710 (1062)  $50(7)$ ,  $10-110(1203)$ 

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  $\frac{1}{2}$  combryos comparisons and synapses. An additional observers and synapses of  $\mathbb{R}^n$ 

William J. Dreyer\* Division of Biology, California Institute of Technology, Pasadena, CA 91125 *William J. Dreyer\** cell surface code for assembling embryos code for assembling embryos code for assembling embryos code for assembling embryos code for a series of a se

**specificity of cell migration and tissue assembly that occurs** 

WJ. Dreyer PNAS 95:9072-9077 (1998) to rule out the presence of either chemical or electrical selectivity in favor of a **ABSTRACT Recent evidence emerging from several labo**aid in the complex control of the expression of the express





 $\mathbb{R} \rightarrow \mathbb{R}$  DE FRANCE Thomas LECUIT 2024-2025



developmental patterning from several and fiber systems in the multiple ratories, integrated with new data obtained by searching the genome databases, suggests that the area code hypothesis pro**basis, particularly in the oriental specificity of cell migration and tissue assembly that occurs** ultroughout embryogenesis. The area code hypothesis proposes<br>that cells assemble organisms, including their brains and neryous systems, with the aid of a molecular-addressing code that tunctions much like the country, area, regional, and local por-<br>tions of the telephone dialing system. The complexity of the **and partly behavior in 1939 and predicted highly specific final participate in 1939** a form of chemical selection of entire organisms is so enormous that we assume that the code must nake combinatorial use of members of large multigene families.  $\frac{3}{2}$  in various regions of the embryo, thus greatly reducing the total number of genes required. We present the hypothesis that 10um vomeronasal receptors fulfill the criteria proposed for area code molecules and could serve as the last dights in such a code. We<br>discuss our evidence indicating that receptors of these families **beginning** are expressed in many parts of developing embryos and suggest that they jaily a high interesting to the car receipting into the three states. ABSTRACT Recent evidence emerging from several laboechome databases, suggests date the architecture replanting the remarkable<br>vides a good heuristic model for explaining the remarkable throughout embryogenesis. The area code hypothesis proposes functions much like the country, area, regional, and local porand as a role more determined and prefix and prefix and prefix completely of the information required to code cells for the construction of entire Such a system would reuse the same receptors as molecular digits members of the very large families of olfactory receptors and voincronasar receptors runnt the criteria proposed for area code.<br> **molecules and could serve as the last digits in such a code.** We are expressed in many parts of developing embryos and suggest<br>that they play a key functional role in cell recognition and

 $\sigma$  bibliographic information. Large numbers of  $\mathcal{L}$ 

- **Cell**<br>• 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)



 $\boldsymbol{\mathcal{C}}$ 

Strong No binding

- $\bullet$  Small families with promiscuous binding for subtype (A) ELISA-based screening method (see text). recognitions. Ex: DIP/Dpr
- $\bullet$  Combinatorial expression. Provides general address code.
- include Dprs, DIPs, Syg1/Syg2-related proteins, Down syn-• Reuse: to minimise size of families.

Reviewed in Sanes JR, Zipursky SL. *Cell*. 181(3):536-556 (2020) with one of six transmetric  $\mathcal{L}$ 



Thomas LECUIT 2024-2025

# 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<sup>1</sup>, Abhishek Kulkarni<sup>1</sup>, Max von Kleist<sup>2\*</sup>, P. Robin Hiesinger<sup>1\*</sup>

 $\frac{1}{\sqrt{2}}$ COLLÈGE DE FRANCE (R cells 1 to 6, herein referred to as R1–6) in  $\frac{1}{\sqrt{2}}$ 

- and axes defined by Arm labeling, the arrange- $\bullet$  Growth cones show biased stochastic search with  $\mathcal{L}$  since  $\mathcal{L}$ . Furthermore, labeling of only  $\mathcal{L}$ \$ignificant overlap typic bundle position (21, 28), revealed correct
- irowth cones search is  $\bullet$  Growth cones search is independent of target cell
- alf-oraznisad filopodi  $\bullet$  §elf-organised filopodia meshwork based on siabhour intoractions heighbour interactions



Thomas LECUIT 2024-2025

 $\frac{1}{2}$ 



F). Hence, the initial neural superposition pattern is established by P35 with  $E.$  Ag1 et  $\epsilon$  data suggest that as  $\epsilon$ 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

- 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.



### Some features of biological encoding properly call that an aperiodic crystal or solid and express our

- 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  $372,529,029,846,191,405.$ 



Thomas LECUIT 2024-2025