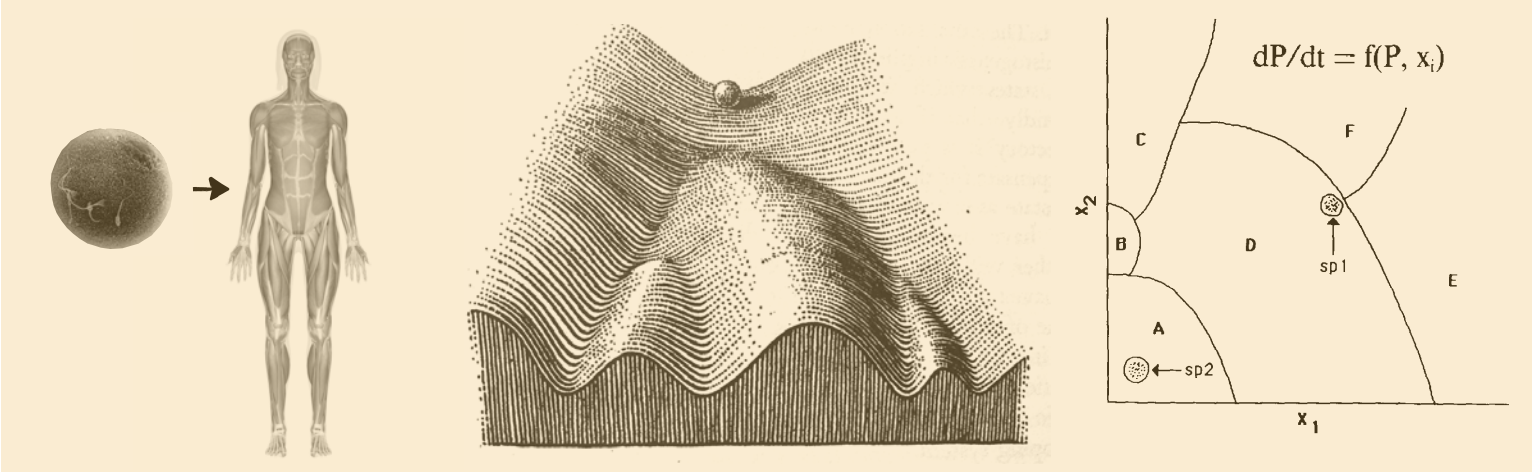
What is the biological information?



Course 1: Introduction –

What representations for the genome?

Thomas Lecuit

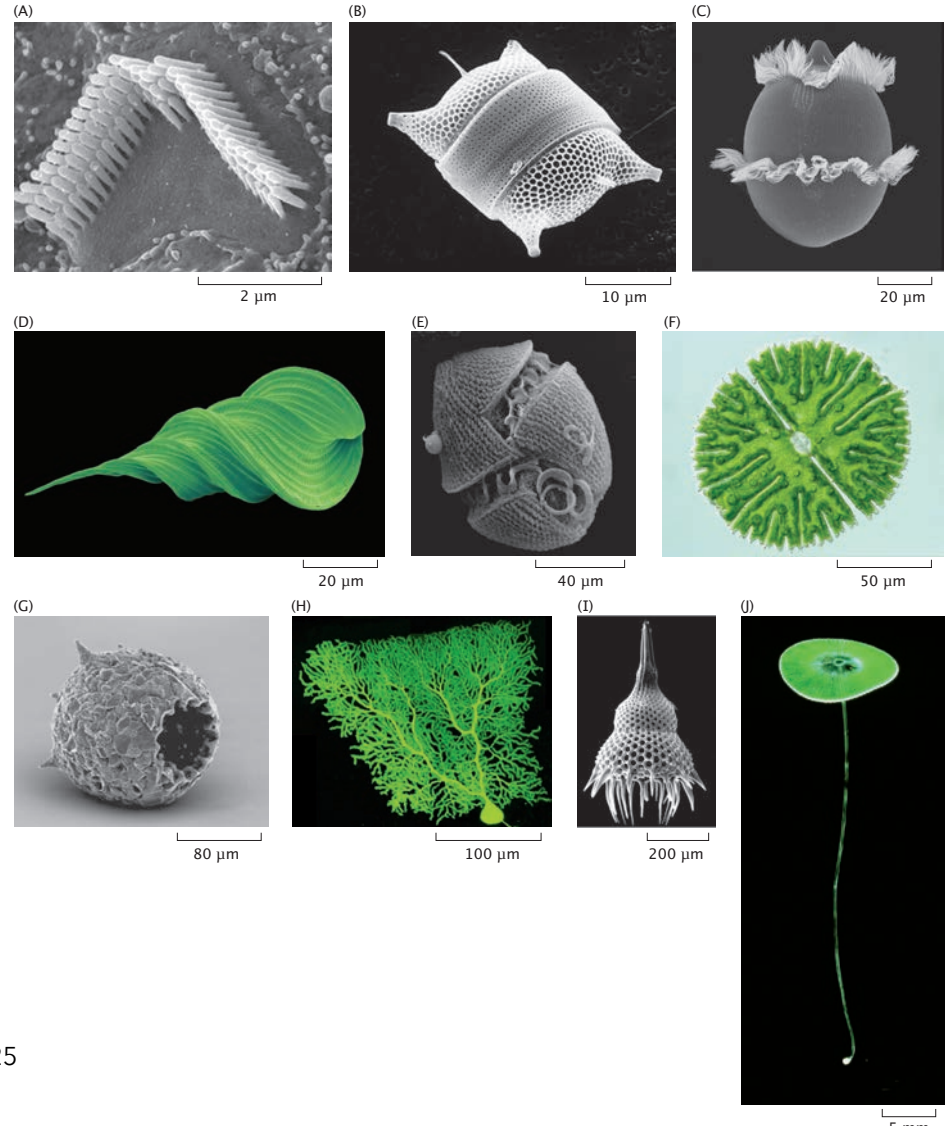
chaire: Dynamiques du vivant

COLLÈGE DE FRANCE 1530

Forms are extremely diverse…

External forms are extremely diverse…

What accounts for dynamical internal and external organisation of cells and organisms?

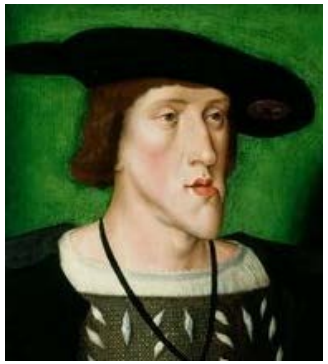


Yet forms are *similar* from parents to progeny

– Dogs make dogs, birds make birds, human make human …

- Stability of forms across generations

Q: What underlies the stability of forms following replication/reproduction?



Charles Quint (1530)

Philippe IV (1630)

Charles II (1685)

Erwin Schrödinger *What is Life?* 1948

10^7 years

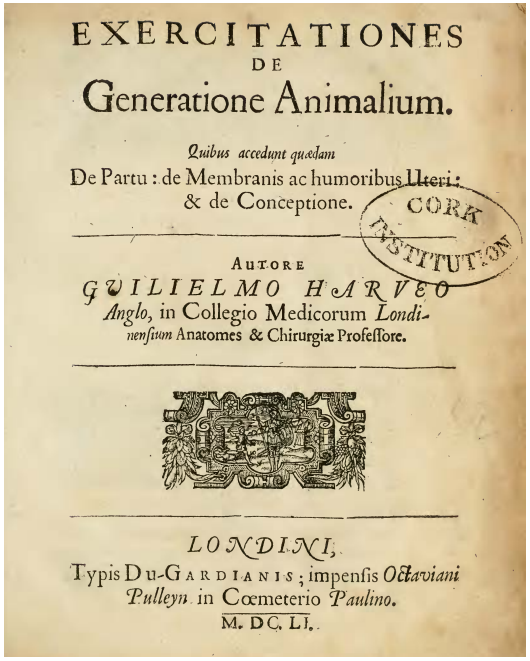
Plan

1. The egg as a « compressed information » state
2. The genome as a carrier of developmental information
3. Metaphors for the genome: blueprint, code script, program, etc.
4. Properties of Genotype to Phenotype mapping
5. Low dimensional representations

Ex ovo omnia

William Harvey
(1578-1657)

All organisms stem from eggs
Against model of « spontaneous generation »



EXERCITATIONES
DE
Generatione Animalium.

Quibus accedunt quædam
De Partu : de Membranis ac humoribus Uteri :
& de Conceptione.

Autore
GVILIELMO HARVEO
Anglo, in Collegio Medicorum *Londinensium* Anatomes & Chirurgiæ Professore.

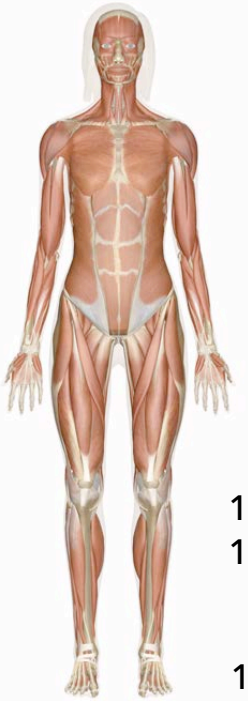
LONDINI,
Typis Du-Gardianis ; impensis *Octaviani Pulleyn* in Cœmeterio *Paulino*.
M. DC. LI.

Heredity - Principle of similarity between parents and offspring

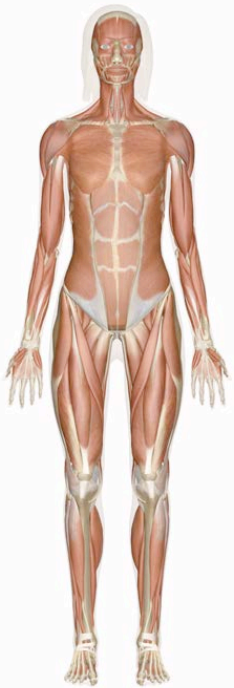
- Organisms emerge from a single cell, the fertilised egg.

Voyez-vous cet œuf? C'est avec cela qu'on renverse toutes les écoles de théologie et tous les temples de la terre.

DIDEROT
Entretien avec d'Alembert.



D. Phillips/SCIENCE PHOTO LIBRARY



1-2m
60-80 kg

400 cell types
30.000.000.000.000 cells

10^4/1
10^{10}/1

10^{13}/1

0.12 mm
1µg

1 cell

1-2m
60-80 kg

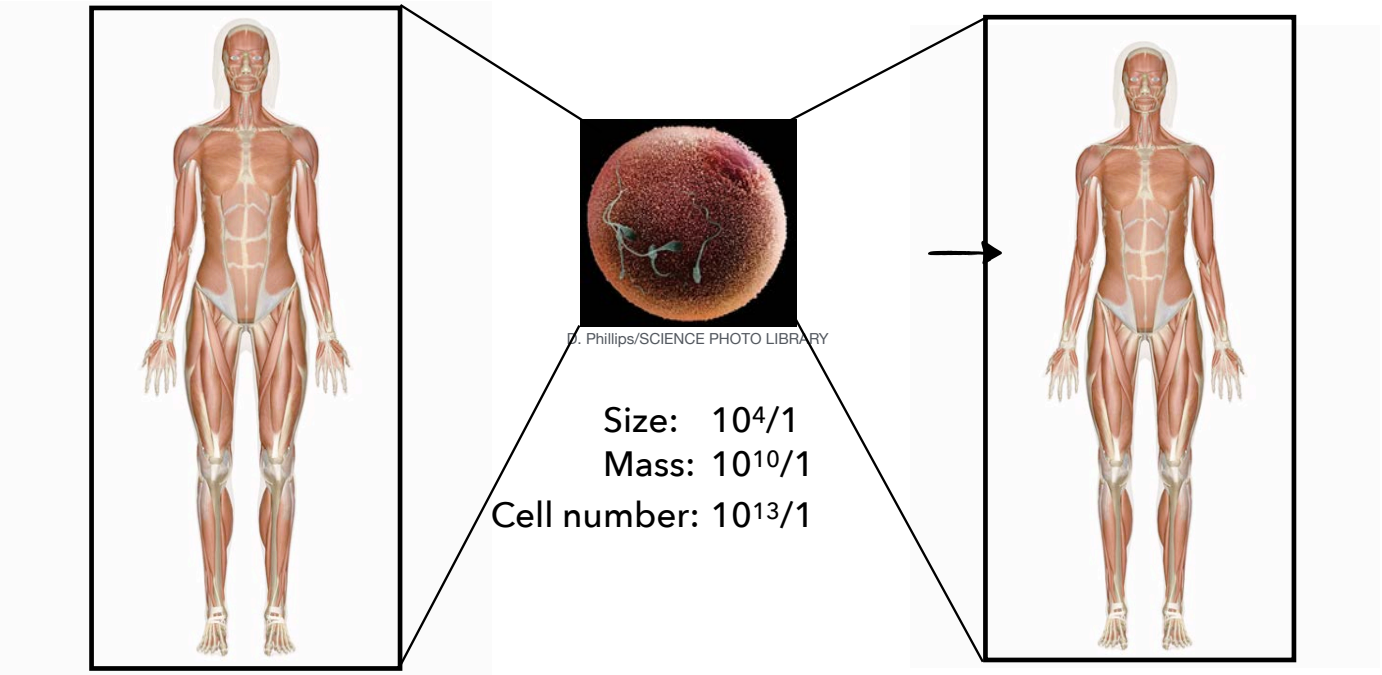
400 cell types
30.000.000.000.000 cells

The egg as a « compressed state » of the organism

- **The complexity of an adult is seemingly compressed/represented** in a single cell

ie. the egg contains all the information needed to rebuild a new organism

- **Consider information as the set of instructions required for this process**



- Questions:
  - What is this information content?
  - What is the *representation* of the future embryo/adult in the egg?

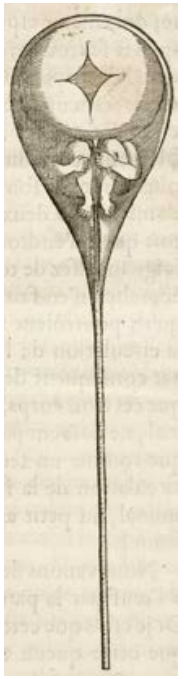
(Paris *intra muros*: ~ 10kmx10km, ~2.000.000 inhabitants, or ~10^{11}g)
Similar compression: size: ~1mx1m, ~10g…)

The egg as a « compressed state » of the organism

- **Preformationism: The *homunculus* or animalcules**

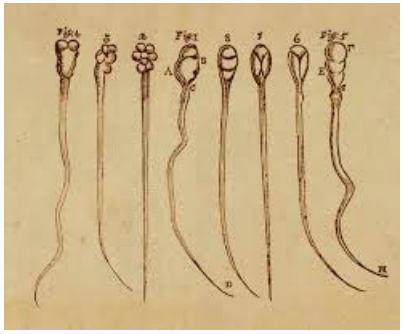
A. van Leeuwenhoek, Hartsoeker, Malebranche

A miniature version of an organism in the egg/sperm
Animal development is an *unfolding* process
Forms emerged at origin of all living forms (creationism)
Theory of « emboitement des germes » (Malebranche)
(embryos contained *ad infinitum)*



Nicolaas Hartsoeker

A. van Leeuwenhoek



- **Epigenesis:**

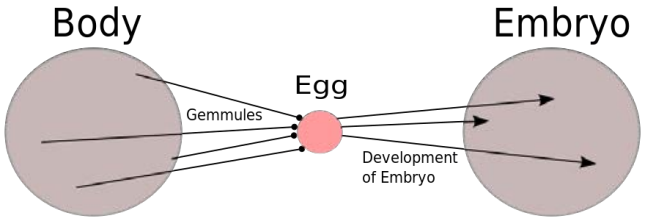
Graduel elaboration of form
Mecanism for the construction of the embryo from an egg (Development)

William Harvey

Models of the germ line

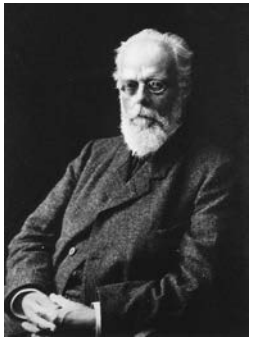
- Pangenesis
  The soma contributes to the constitution of the egg

Maupertuis (particles)
Buffon (organic molecules)
Lamarck
Darwin (gemmules)
De Vries (intracellular pangenesis)

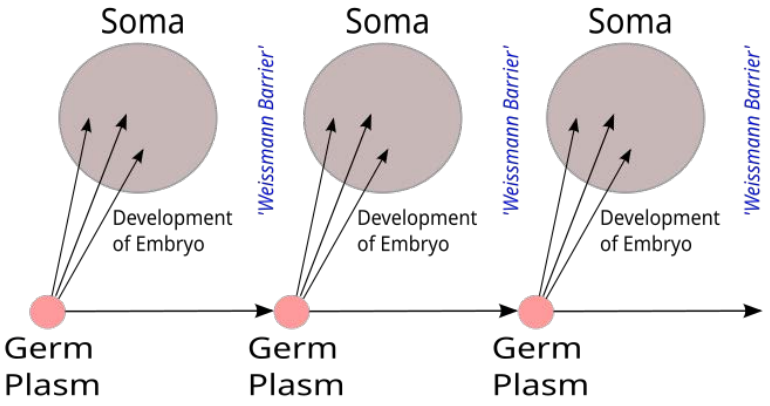


- Weismann barrier
  The soma does not affect the germ line

A. Weismann: germ plasm theory
Immortal germ cells
Mortal soma



"Das Keimplasma: eine Theorie der Vererbung" (1892)

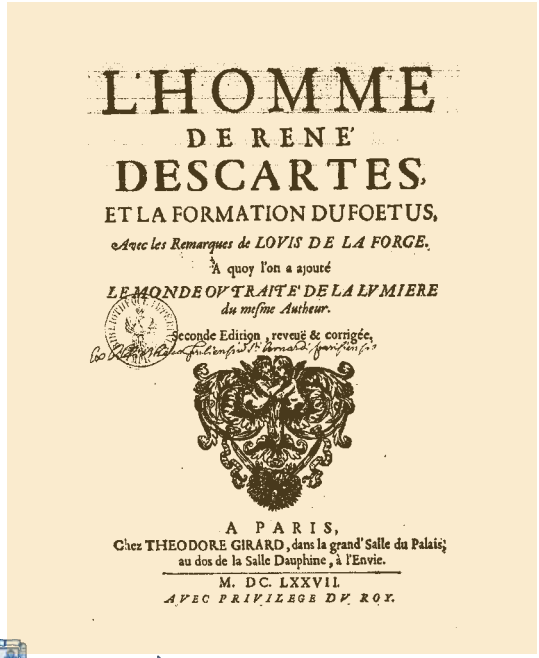


Descartes - Animal as a machine

- Animals are mere machines or automaton, powered by heat

R. Descartes (1596-1650)

Discours de la Méthode (5ème partie), 1637, *Traité de l'homme*, 1662



L'HOMME
DE RENE'
DESCARTES
ET LA FORMATION DU FOETUS,
Avec les Remarques de LOVIS DE LA FORGE.
A quoy l'on a ajouté
LE MONDE OV TRAITE' DE LA LVMIERE
du mesme Autheur.
Seconde Edition, reveuë & corrigée.
A PARIS,
Chez THEODORE GIRARD, dans la grand' Salle du Palais;
au dos de la Salle Dauphine, à l'Envie.
M. DC. LXXVII.
AVEC PRIVILEGE DV ROY.

CVI. Que toutes les fonctiõs qui luy ont eſté attribuées ſont des ſuittes de la diſpoſition de ſes organes.

Je deſire que vous conſideriez aprés cela, que toutes les fonctions que j'ay attribuées à cette Machine, comme la digeſtion des viandes, le battement du cœur & des arteres, la nourriture & la croiſſance des membres, la reſpiration, la veille & le ſommeil; la reception de la lumiere, des ſons, des odeurs, des gouſts, de la chaleur, & de telles autres qualitez, dans les organes des ſens exterieurs; l'impreſſion de leurs idées dans l'organe du ſens commun & de l'imagination; la retention ou l'empreinte de ces idées dans la Memoire; les mouvemens interieurs des Appetits, & des Paſſions; Et enfin les mouvemens exterieurs de tous les Membres, qui ſuivent ſi à propos, tant des actions des objets qui ſe preſentent aux ſens, que des paſſions, & des impreſſions qui ſe rencontrẽt dans la Memoire, qu'ils imitent le plus parfaitement qu'il eſt poſſible ceux d'vn vray homme; Je deſire, dis-je, que vous conſideriez que ces fonctions ſuivent toutes naturellement en cette Machine, de la ſeule diſpoſition de ſes organes; ne plus ne moins que font les mouvemens d'vne horloge, ou autre automate, de celle de ſes contrepoids & de ſes roües; En ſorte qu'il ne faut point à leur occaſion concevoir en elle aucune autre Ame vegetative, ny ſenſitive, ny aucun autre principe de mouvement & de vie, que ſon ſang & ſes Eſprits agitez par la chaleur du feu qui brûle continuellement dans ſon cœur, & qui n'eſt point d'autre Nature que tous les feux qui ſont dans les Corps Inanimez.

Source: *Gallica, BNF*

VAI. Sommaire des choſes qu'il doit contenir.

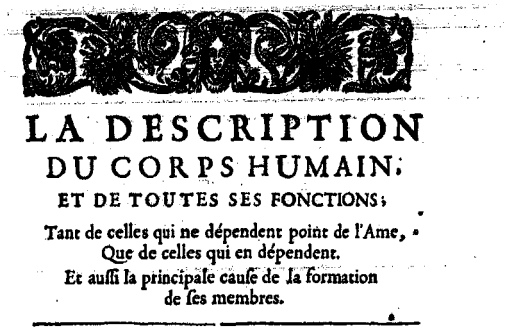
Et afin qu'on ait d'abord vne generale notion de toute la Machine que j'ay à décrire; Ie diray icy que c'eſt la chaleur qu'elle a dans le Cœur, qui eſt comme le grand Reſſort, & le Principe de tous les mouvemens qui ſont en elle

Descartes - Animal as a machine

- Animals are similar to clockworks, albeit with smaller components.
- Source of motion is *heat*.
- Efficient causality and determinism

« Je ne reconnais aucune différence entre les machines que font les artisans et les divers corps que la nature seule compose. (…). Et il est certain que toutes les règles des Mécaniques appartiennent à la Physique, en sorte que toutes les choses qui sont artificielles, sont avec cela naturelles. Car, par exemple, lorsqu'une montre marque les heures par le moyen des roues dont elle est faite, cela ne lui est pas moins naturel qu'il est à un arbre… de produire ses fruits. »

Principes de la Philosophie , 1644.



LA DESCRIPTION
DU CORPS HUMAIN:
ET DE TOUTES SES FONCTIONS;
Tant de celles qui ne dépendent point de l'Ame,
Que de celles qui en dépendent.
Et auſſi la principale cauſe de la formation
de ſes membres.

LXVI. Que de la connoiſſance des parties de la ſemence on pourroit déduire la figure & la conformation de tous les membres.

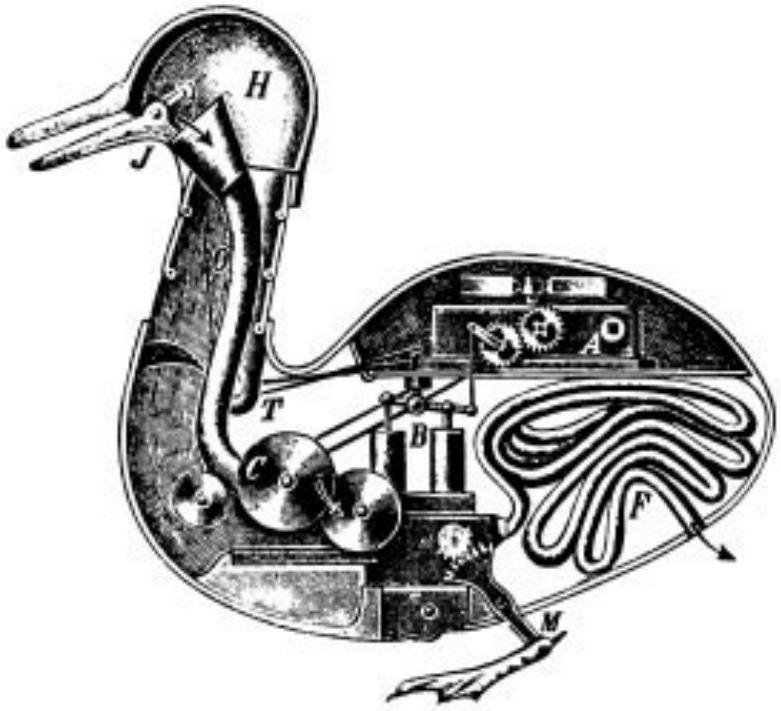
Or dautant que les petits filets dont les parties ſolides ſont compoſées, ſe détournent, ſe plient, & s'entrelacent en diverſes façons, ſuivant les divers cours des matieres fluides & ſubtiles qui les environnent, & ſuivant la figure des lieux où ils ſe rencontrét, ſi on connoiſſoit bien quelles ſont toutes les parties de la ſemence de quelque eſpece d'Animal en particulier, par exemple *de l'homme*, on pourroit déduire de cela ſeul, par des raiſons entierement Mathematiques & certaines, toute la figure & conformation de chacun de ſes membres

p. 146

Animal as a machine

The duck automaton, *Le canard digérateur*

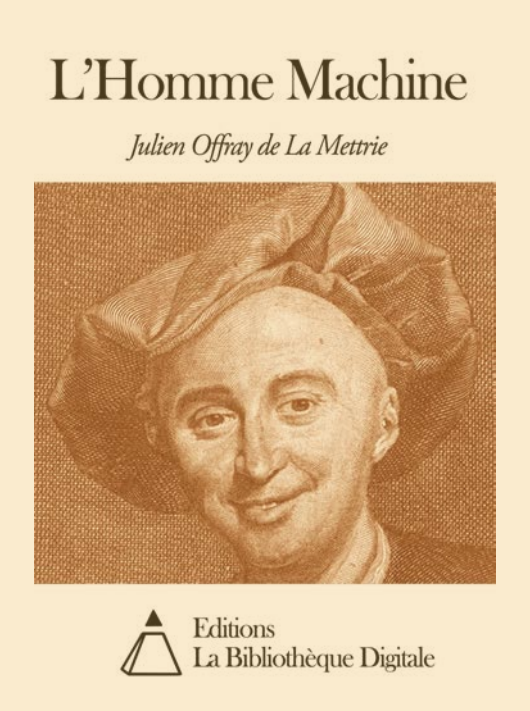
Vaucanson, 1734



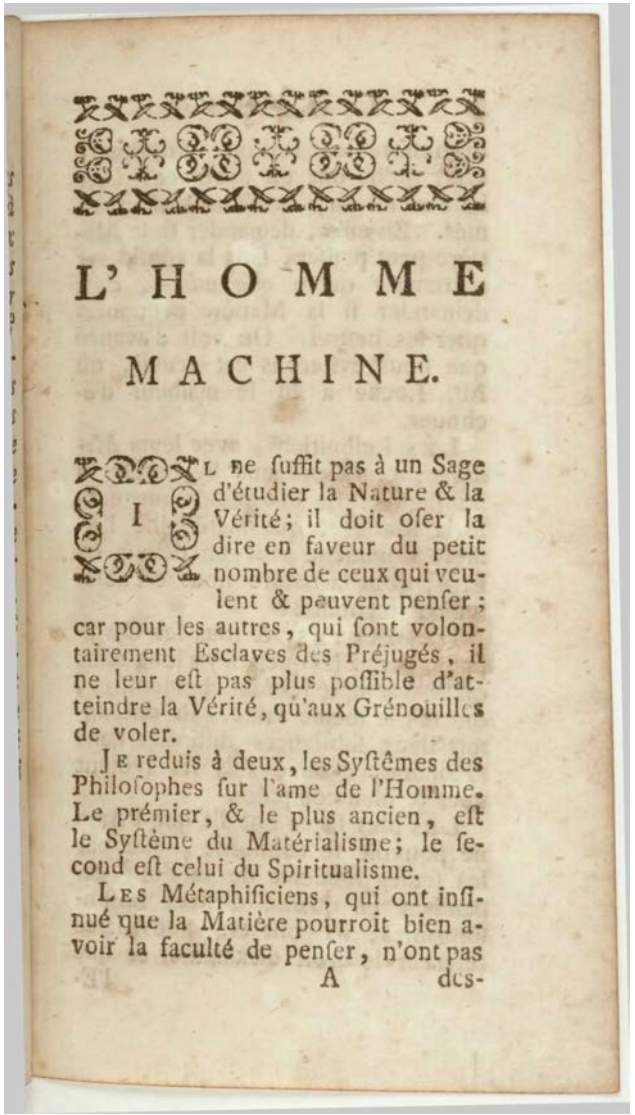
Voltaire:
« Le hardi Vaucanson, rival de Prométhée
Semblait, de la nature imitant les ressorts,
Prendre le feu des cieux pour animer les corps. »

La Mettrie - Man as a machine

- Extends to humans the automaton/machine concept
- Materialism: only one substance, even for man (mind)
- Deterministic framework



(1709-1751)


L' HOMME
MACHINE.

IL ne ſuffit pas à un Sage d'étudier la Nature & la Vérité; il doit oſer la dire en faveur du petit nombre de ceux qui veulent & peuvent penſer; car pour les autres, qui ſont volontairement Eſclaves des Préjugés, il ne leur eſt pas plus poſſible d'atteindre la Vérité, qu'aux Grénouilles de voler.

JE reduis à deux, les Syſtêmes des Philoſophes ſur l'ame de l'Homme. Le prémier, & le plus ancien, eſt le Syſtême du Matérialisme; le ſecond eſt celui du Spiritualisme.

LES Métaphiſiciens, qui ont inſinué que la Matière pourroit bien avoir la faculté de penſer, n'ont pas

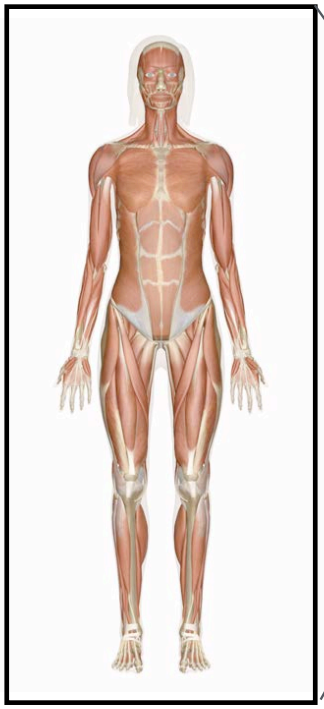
A des-

1748

The egg as a« compressed information state » of the organism

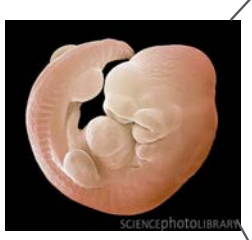
compression

decompression algorithm?

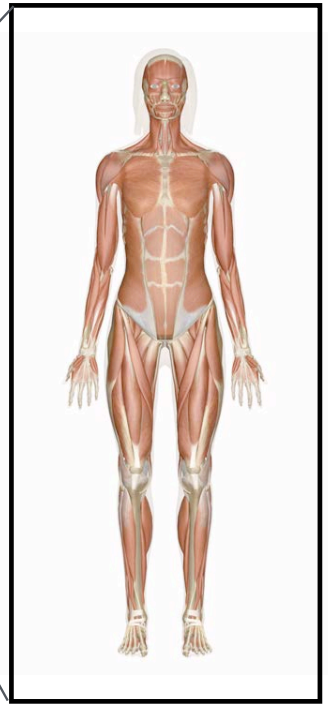


D. Phillips/SCIENCE PHOTO LIBRARY

Size: $10^4/1$
Mass: $10^{10}/1$
Cell number: $10^{13}/1$



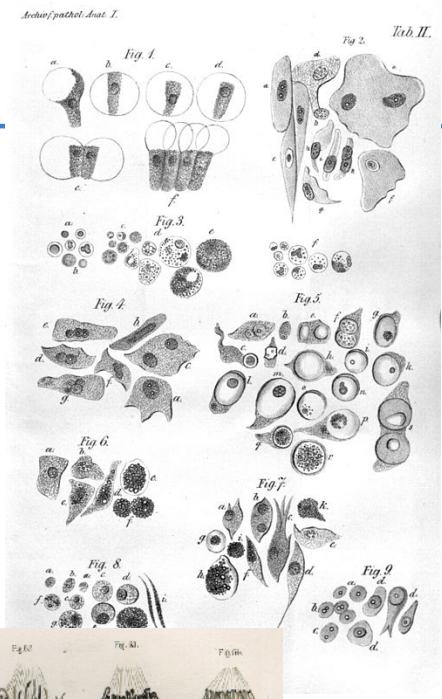
development



Plan

1. The egg as a « compressed information » state
2. The genome as a carrier of developmental information
3. Metaphors for the genome: blueprint, code script, program, etc.
4. Properties of Genotype to Phenotype mapping
5. Low dimensional representations

What in the egg is a carrier of heredity?



- **Cell theory:**

–T. Schwann (1839), the cell as a fundamental unit and building block

–R. Virchow (1855), « *omnis cellula e cellula* » , every cell arises from a dividing cell (initially discovered by R. Remak)

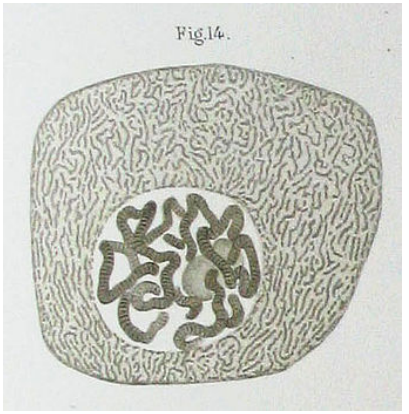
–W. Flemming (1882), cell division and nuclei: « *omnes nucleus e nucleo* »

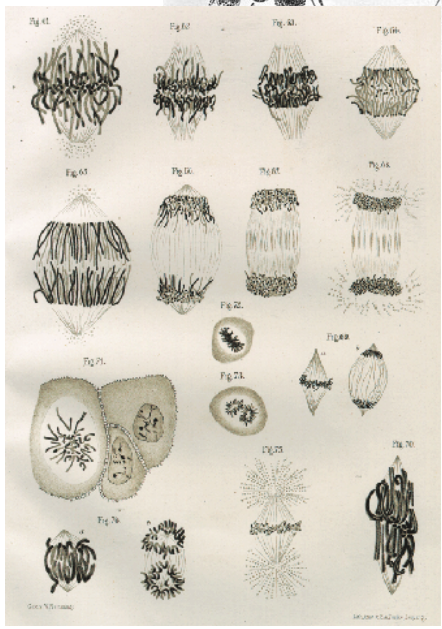
Characterizes the structure called chromatin in nuclei

Later referred to as chromosomes by HW. Waldeyer

Zellsubstanz, Kern und Zelltheilung (1882)

- **Chromosomes**
  - Chromosomic theory of heredity: Sutton and Boveri (1902-1905)
  - Role in development: T. Boveri.

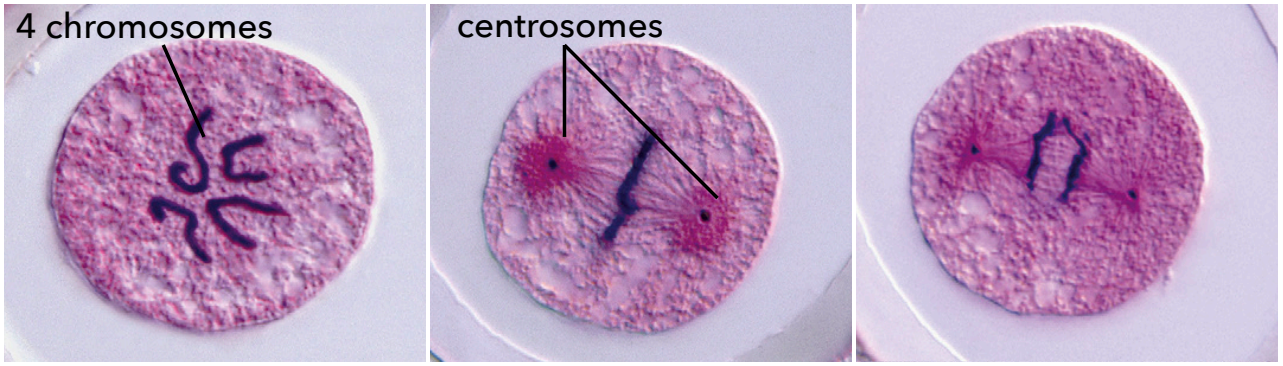




What in the egg is a carrier of heredity?

- Discovery of *chromosome continuity and individuality* (T. Boveri)

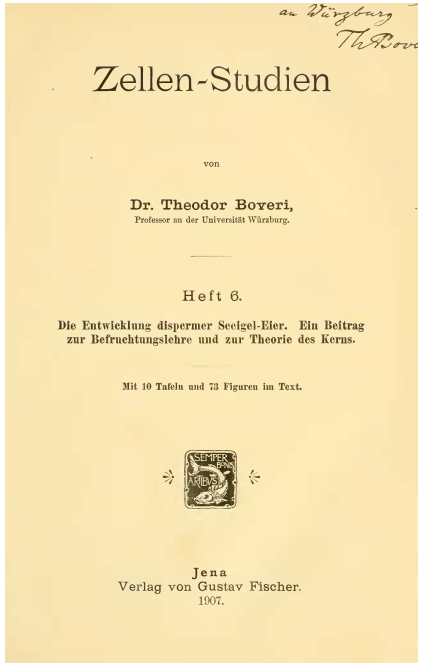
– Discovery of centrosomes, that attach to « chromatic elements/loops », ie. chromosomes (1887-90)
– one chromosome is connected by 2 centrosomes
– key role in chromosome segregation
– Boveri tracked the position of chromosomes (4) and showed they remained in the same position after division in daughter cells.
– Hence **chromosome continuity and individuality**: chromosomes have a stable identity



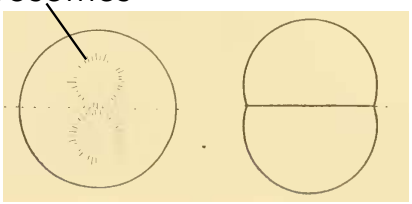
Ascaris egg (nematode)

The chromosome as a determinant of organism form

Theodor Boveri
(1862-1915)



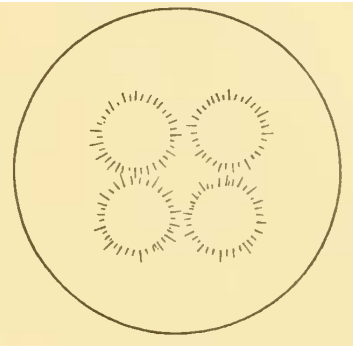
2 centrosomes



System:
dispermic eggs form tetrasters
2 sperms + 1oocyte:

- 3 sets of chromosomes (3x36)
- 4 centrosomes

>> 4 cells form at once, with expected anomalies in chromosome segregation



Development of dispermic Sea Urchin eggs. Contribution to the study of fertilisation and to the theory of the nucleus, Theodor BOVERI, 1907

The chromosome as a determinant of organism form

- Tetracentric embryos

- 54 (3x18) chromosomes connect to a pair of centrosomes and do so randomly among the 4 centrosomes.
- **Provides a means to modify the chromosome content of individual blastomeres and hence study their function.**

1 cell
54 chr
(3 haploid genomes)

1 cell
108 chr
(3 diploid 2N)

4 cells

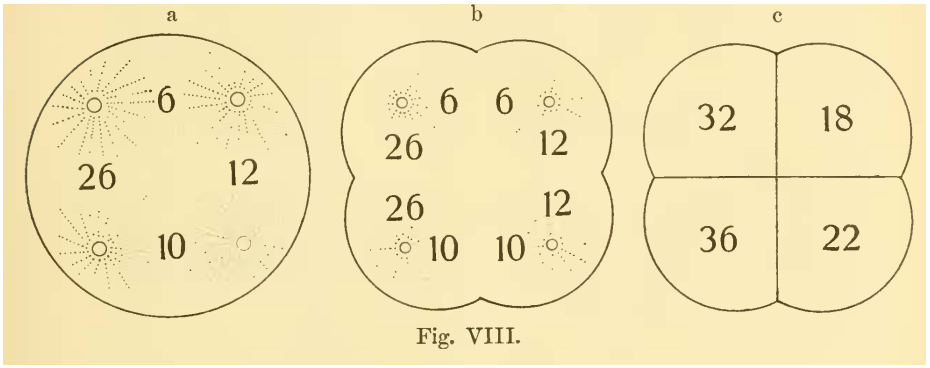
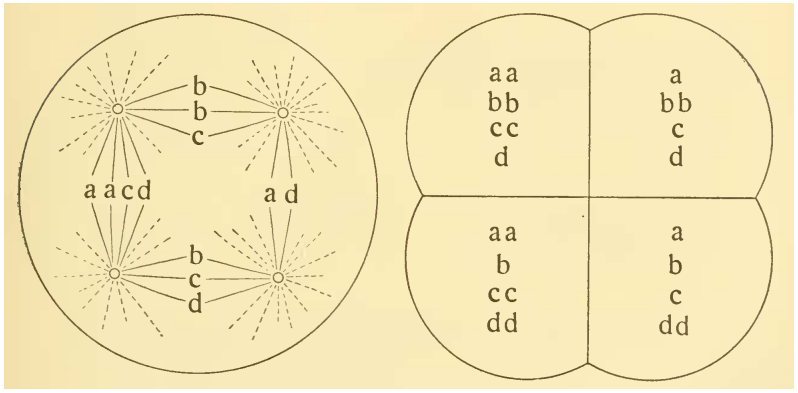


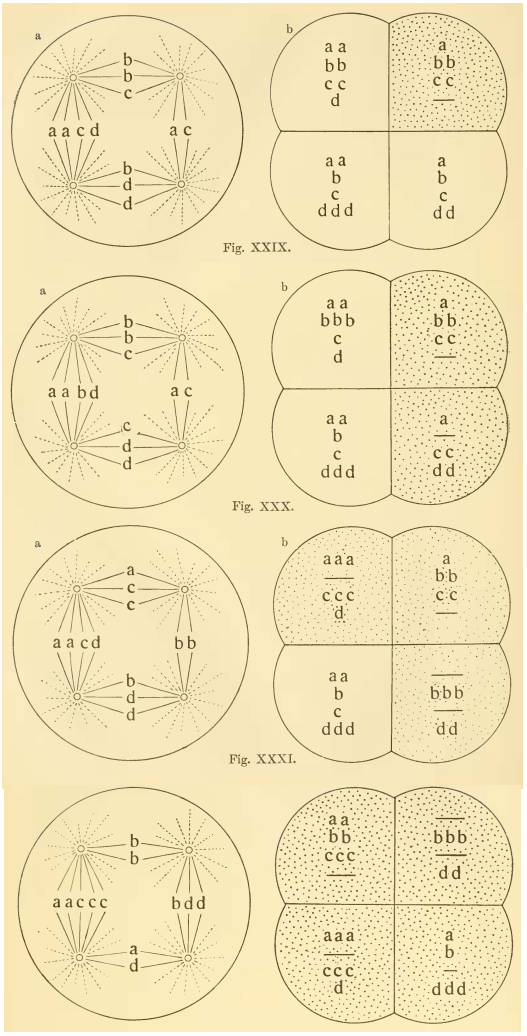
Fig. VIII.



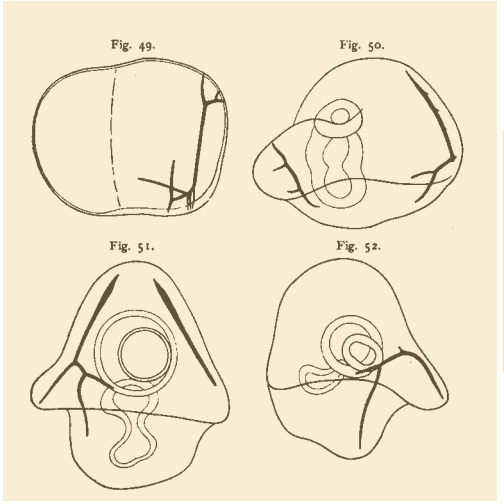
Development of dispermic Sea Urchin eggs. Contribution to the study of fertilisation and to the theory of the nucleus, Theodor BOVERI, 1907

The chromosome as a determinant of organism form

- **Predictions** of hypothetical chromosomal segregation patterns in tetraster eggs
- Resonated with hereditary patterns from Mendel, rediscovered in ~1900.



- **Boveri analysed the fate of developing sea urchin embryos**
- Embryos showed various defects/phenotypes.
- *Hypothesis: reflects defects in chromosome segregation*
- To test this further, Boveri split/disassembled the tetrasters in 4 blastomeres and studied their fate (H. Driesch)
- Observation of a variety of fates from each blastomere



1) Zerlegungsversuche

ganz normal	0	Proz.
3/4 normal	4,5	„
2/4 „	4,5	„
1/4 „	54,5	„
ganz pathologisch	36,5	„

The chromosome as a determinant of organism form

- **Quantitative approach by comparison of tetraster and triaster eggs**

  (To circumvent the fact that he could not track directly chromosome content per cell/phenotype)

- In tetrasters: 108 chromosomes to split randomly among 4: average 27/cell
- In triasters: 108 chromosomes among 3: 36/cell which is the correct number.

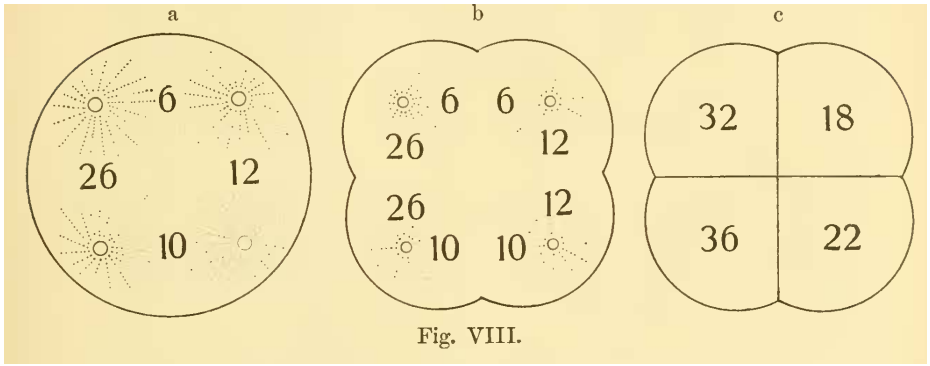


Fig. VIII.

- **Boveri observed more normal looking embryos in triasters, consistent with more frequent expected normal segregation of chromosomes.**

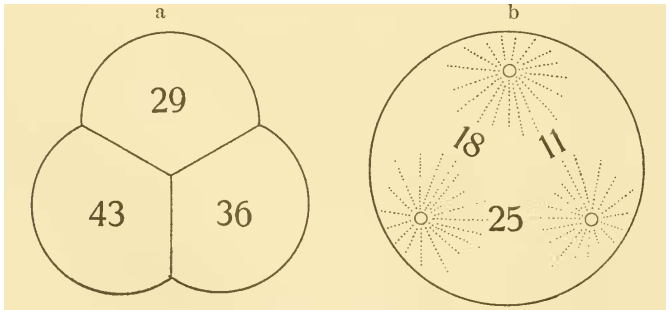


Fig. XLVIII.

Development of dispermic Sea Urchin eggs. Contribution to the study of fertilisation and to the theory of the nucleus, Theodor BOVERI, 1907

The chromosome as a determinant of organism form

The 1st Genotype to Phenotype mapping

- « Simulation » of the biological experiment:

  3 sets of 18 labelled wood beads. Mixed and poured on a plate with 4 or 3 quadrants and looked at the statistics of segregation

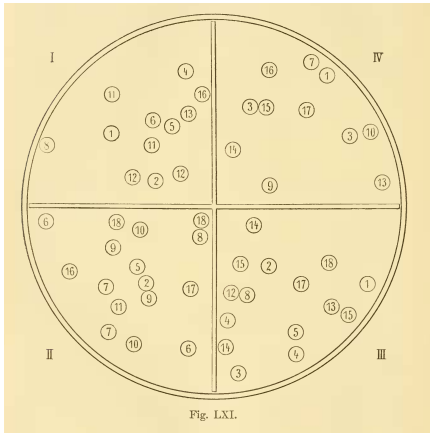


Fig. LXI.

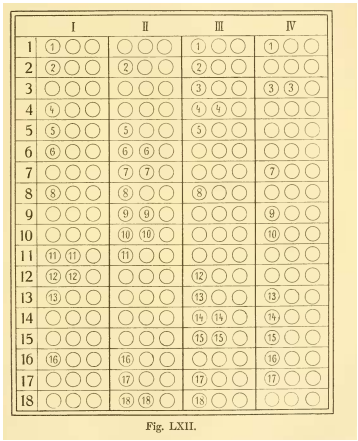


Fig. LXII.

Blastomere A kombiniert aus I und II	Blastomere B kombiniert aus II und III	Blastomere C kombiniert aus III und IV	Blastomere D kombiniert aus IV und I
1	1	1, 1	1, 1
2, 2	2, 2	2	2
—	3	3, 3, 3	3, 3
4	4, 4	4, 4	4
5, 5	5, 5	5	5
6, 6, 6	6, 6	—	6
7, 7	7, 7	7	7
8, 8	8, 8	8	8
9, 9	9, 9	9	9
10, 10	10, 10	10	10
11, 11, 11	11	—	11, 11
12, 12	12	12	12, 12
13	13	13, 13	13, 13
—	14, 14	14, 14, 14	14
—	15, 15	15, 15, 15	15
16, 16	16	16	16, 16
17	17, 17	17, 17	17
18, 18	18, 18, 18	18	—
28 Stück	31 Stück	26 Stück	23 Stück

Data

A. Dreier.

1) Zerlegungsversuche		2) Nachahmung	
ganz normal	14,4 Proz.	ganz normal	11 Proz.
2/3 normal	22,8 „	2/3 normal	42 „
1/3 „	40 „	1/3 „	36 „
ganz pathologisch	22,8 „	ganz pathologisch	11 „

B. Vierer.

1) Zerlegungsversuche		2) Nachahmung	
ganz normal	0 Proz.	ganz normal	0 Proz.
3/4 normal	4,5 „	3/4 normal	0 „
2/4 „	4,5 „	2/4 „	2 „
1/4 „	54,5 „	1/4 „	34 „
ganz pathologisch	36,5 „	ganz pathologisch	64 „

Simulation

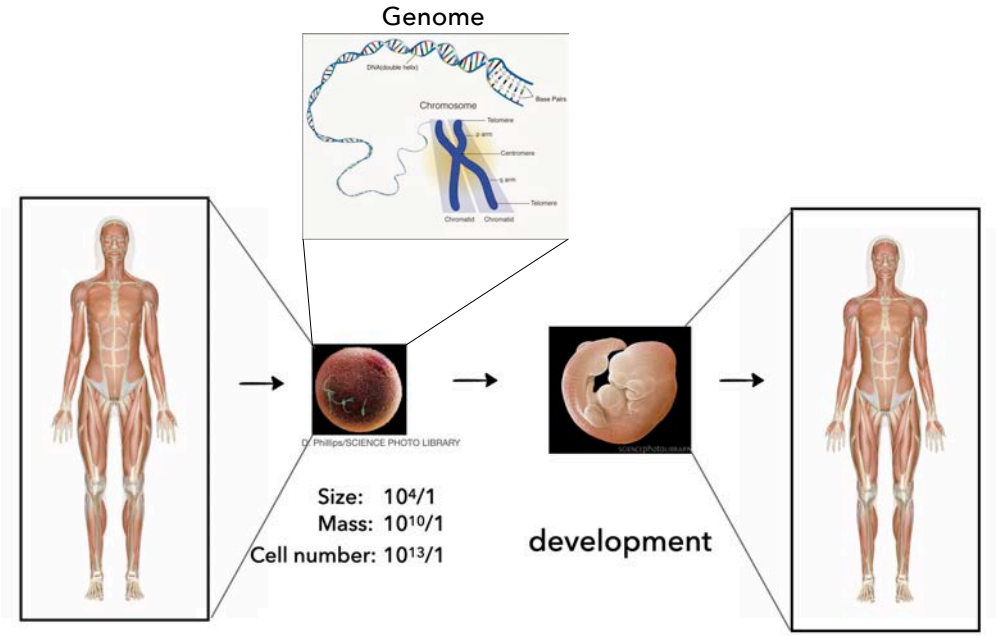
Development of dispermic Sea Urchin eggs. Contribution to the study of fertilisation and to the theory of the nucleus, Theodor BOVERI, 1907

Plan

1. The egg as a « compressed information » state
2. The genome as a carrier of developmental information
3. Metaphors for the genome: blueprint, code script, program, etc.
4. Properties of Genotype to Phenotype mapping
5. Low dimensional representations

Metaphors for the Genome

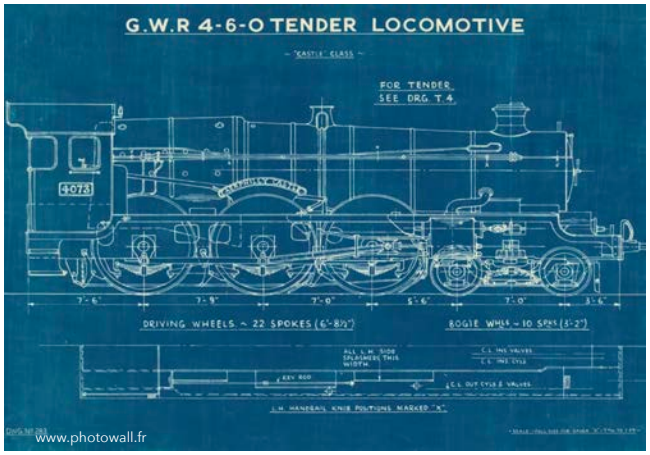
Information bottleneck



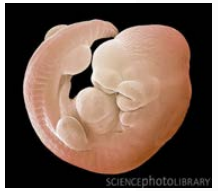
- Blueprint
- Code-script
- Program
- Recipe
- Generative model

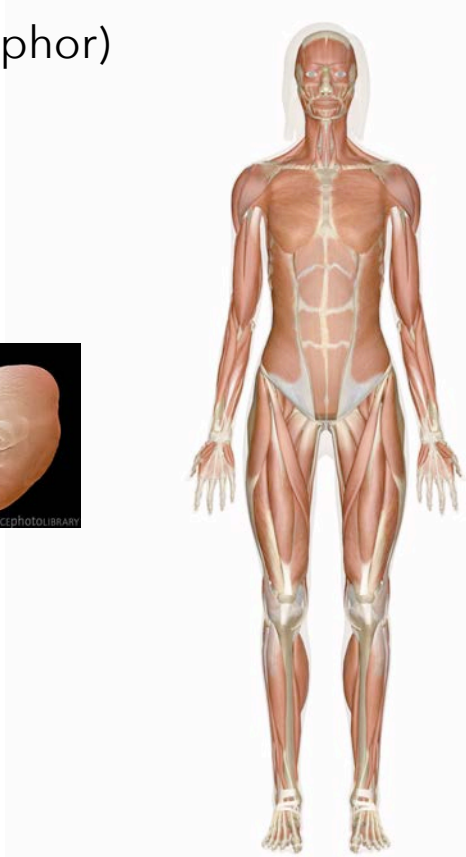
The genome as a blueprint

- **The blueprint metaphor: a preformationist concept**
  - detailed miniaturised plan, pre-specified/determined, that works as a reference
  - Isomorphic to final outcome
  - specifies the size, shape and list of components
  - fully deterministic in all details (cf. Descartes and machine metaphor)
  - does not specify « how to build » the object/organism



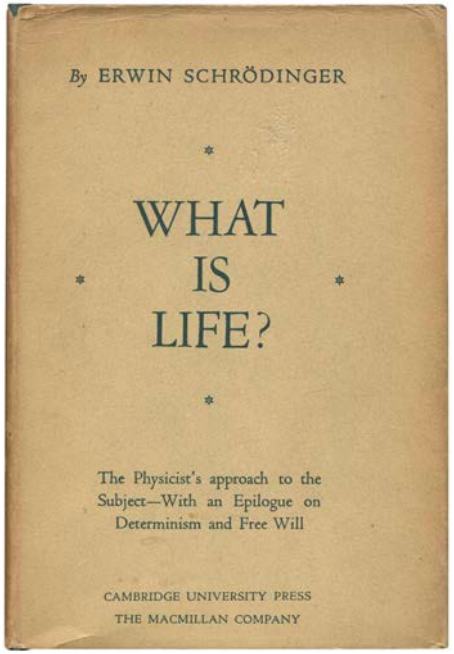
?





The genome as a code-script

E. Schrödinger
(1887-1961)



1944

THE HEREDITARY CODE-SCRIPT (CHROMOSOMES)

Let me use the word 'pattern' of an organism in the sense in which the biologist calls it 'the four-dimensional pattern', meaning not only the structure and functioning of that organism in the adult, or in any other particular stage, but the whole of its ontogenetic development from the fertilized egg cell to the stage of maturity, when the organism begins to reproduce itself. Now, this whole four-dimensional pattern is known to be determined by the structure of that one cell, the fertilized egg. Moreover, we know that it is essentially determined by the structure of only a small part of that cell, its nucleus. This nucleus, in the ordinary 'resting state' of the cell, usually appears as a network of chromatine,[1] distributed over the cell. But in the vitally important processes of cell division (mitosis and meiosis, see below) it is seen to consist of a set of particles, usually fibre-shaped or rod-like, called the chromosomes, which number 8 or 12 or, in man, 48.

It is these chromosomes, or probably only an axial skeleton fibre of what we actually see under the microscope as the chromosome, that contain in some kind of code-script the entire pattern of the individual's future development and of its functioning in the mature state. Every complete set of chromosomes contains the full code; so there are, as a rule, two copies of the latter in the fertilized egg cell, which forms the earliest stage of the future individual.

Rationalist, deterministic view

In calling the structure of the chromosome fibres a code-script we mean that the all-penetrating mind, once conceived by Laplace, to which every causal connection lay immediately open, could tell from their structure whether the egg would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhododendron, a beetle, a mouse or a woman. To which we may add, that the appearances of the egg cells are very often remarkably similar; and even when they are not, as in the case of the comparatively gigantic eggs of birds and reptiles, the difference is not so much in the relevant structures as in the nutritive material which in these cases is added for obvious reasons.

The genome as a code-script

E. Schrödinger
(1887-1961)

- **Compression:**

How can the code-script be so miniature, in a single/2 copies?
–Highly ordered atomic arrangement (« aperiodic crystal »)
–Combinatorial arrangements

- **The *fidelity* conundrum:**

How can a cell encode the high-fidelity information/code-script from which the future organism is built? Expected precision is supposed to scale as $1/N^{1/2}$, with N particles.

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.

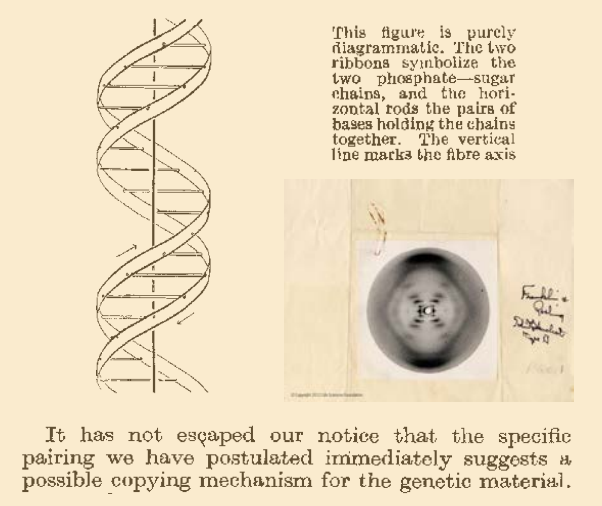
From chromosomes to DNA

- DNA on chromosomes is the carrier of hereditary information

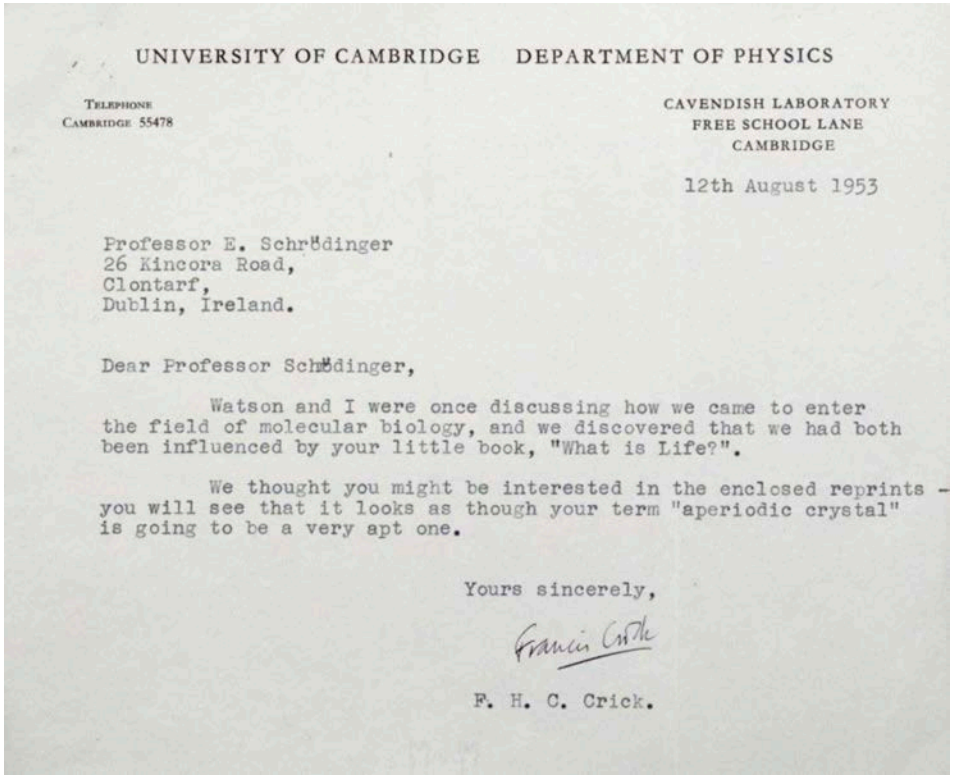
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

April 25, 1953 NATURE



DNA – *Watson and Crick* 1953
& Rosalind Franklin, M. Wilkins



UNIVERSITY OF CAMBRIDGE DEPARTMENT OF PHYSICS

TELEPHONE
CAMBRIDGE 55478

CAVENDISH LABORATORY
FREE SCHOOL LANE
CAMBRIDGE

12th August 1953

Professor E. Schrödinger
26 Kincora Road,
Clontarf,
Dublin, Ireland.

Dear Professor Schrödinger,

Watson and I were once discussing how we came to enter the field of molecular biology, and we discovered that we had both been influenced by your little book, "What is Life?".

We thought you might be interested in the enclosed reprints - you will see that it looks as though your term "aperiodic crystal" is going to be a very apt one.

Yours sincerely,

Francis Crick

F. H. C. Crick.

From chromosomes to DNA

- Characterization of genomic information flow

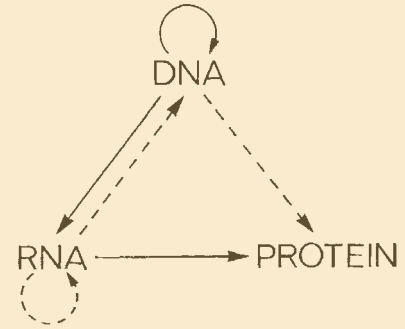
Central Dogma of Molecular Biology

by

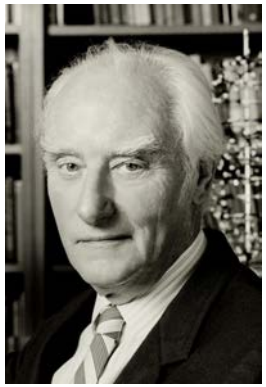
FRANCIS CRICK

MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH

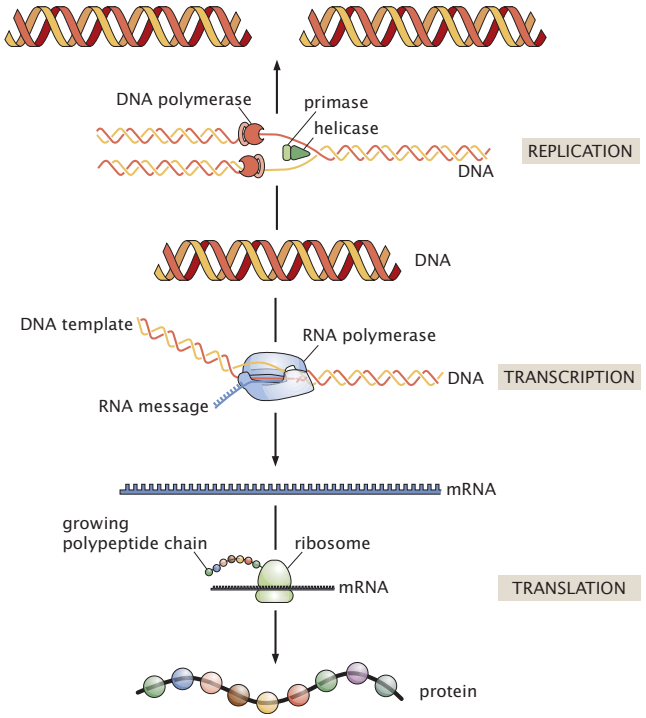
The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.



The central dogma – *Crick* 1970



F. Crick (1916-2004)

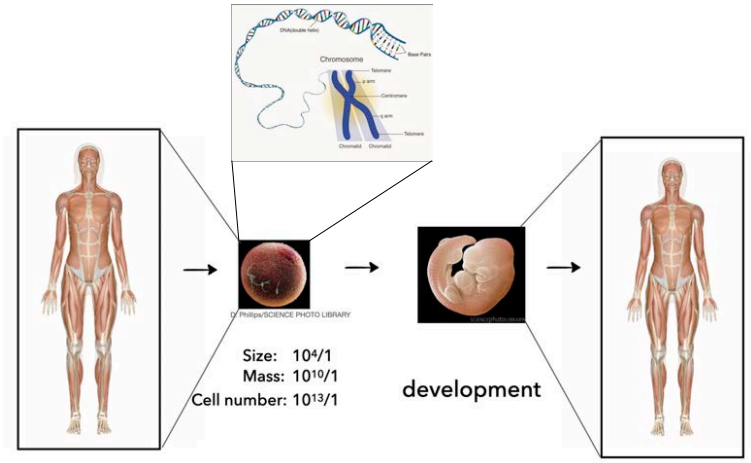


R. Phillips, J. Kondev, J. Thériot & H. Garcia. *Physical Biology of the Cell (Garland Science)* 2012

The genome as a code-script

- Density of information in the genome

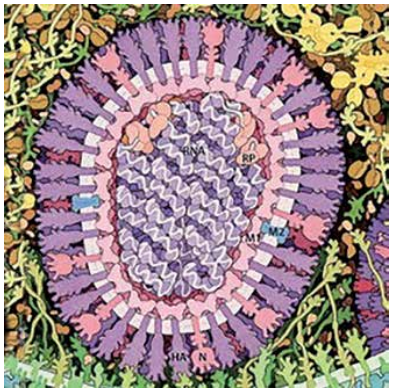
Information bottleneck



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

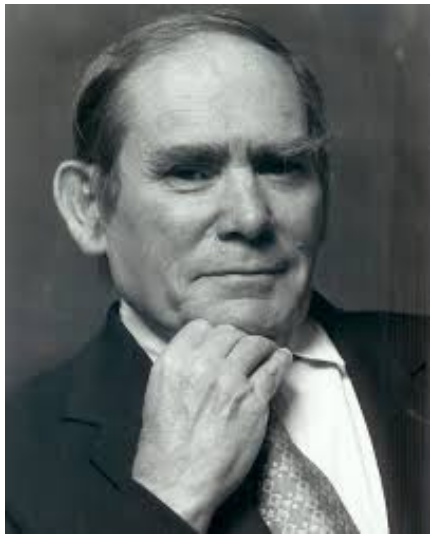


$$\frac{10{,}000 \text{ letters}}{10^5 \text{ nm}^3} \approx 10^{20} \frac{\text{letters}}{\text{cm}^3}$$

Human: ~ 10^9 letters / 10 µm³

The genome as a codescript

- « Schrödinger's fundamental error »



Sydney Brenner (1927-2019)

But the term code-script is, of course, too narrow. The chromosome structures are at the same time instrumental in bringing about the development they foreshadow. They are law-code and executive power – or, to use another simile, they are architect's plan and builder's craft – in one.

Chromosomes are not builders

"Schrödinger says that the chromosomes contain the information to specify the future organism and *the means to execute it* and thats not true. The chromosomes contain the information to specify the future organization and *a description of the future means to implement it, but not the means themselves."*

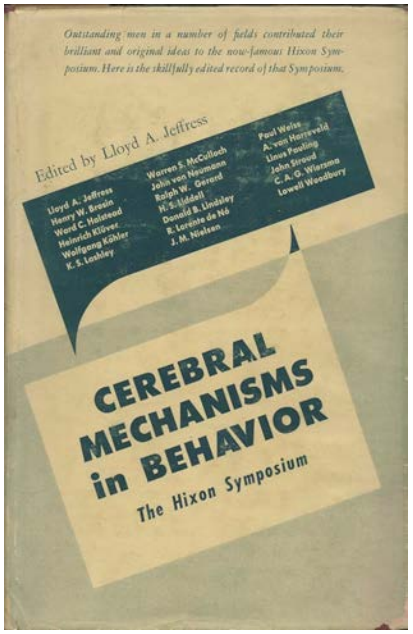
"The code script contains a description of the executive function, not the function itself."

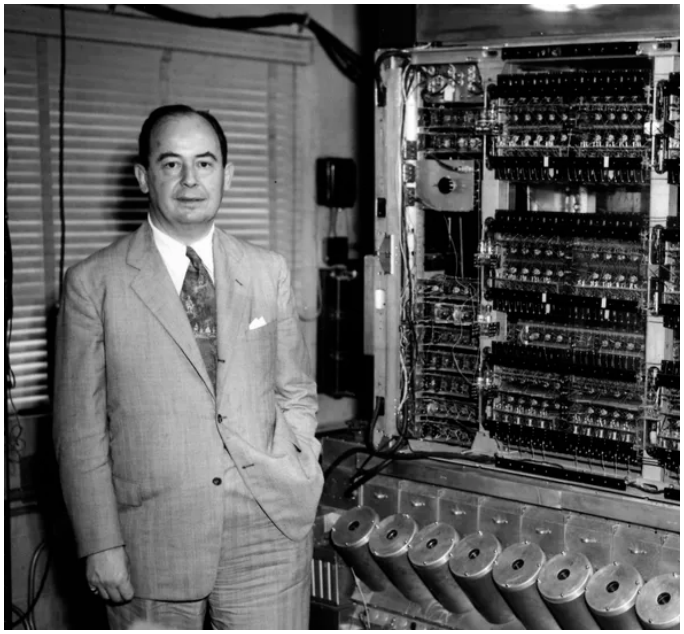
Self-reproducing automata

The General and Logical Theory of Automata

John von Neumann (1903-1957)

conference, 1948. publication,1951





- Established a link between the ability of cells and organisms to self-reproduce and the theory of universal computation in automata/machines developed by Turing (1936).

- *According to this view, Life is intimately linked to computation and information processing*

Von Neumann, J., 1951. In: Jeffress, L.A. (Ed.), *Cerebral Mechanisms of Behavior: The Hixon Symposium.* John Wiley and Sons, New York, pp. 1–41.

Self-reproducing automata

The General and Logical Theory of Automata

John von Neumann (1903-1957)

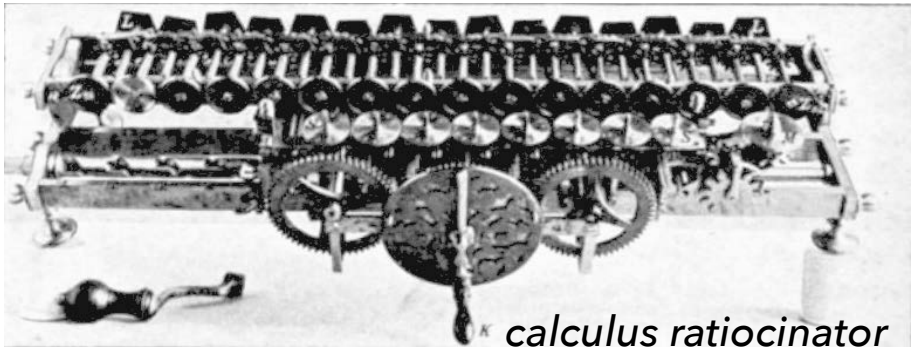
At the confluence of two philosophical heritages

– Living organisms are deterministic machines (Descartes, 1662)
– Logic and rationality (Leibniz,1666): *calculus ratiocinator* using the *characteristica universalis*

A formal language to decide whether a proposition is true or false.
Used in a machine that conducts universal logical calculus, including arithmetic operations

"The history of the modern computing machine goes back to Leibniz and Pascal. Indeed, the general idea of a computing machine is nothing but a mechanization of Leibniz's *calculus ratiocinator*."

Norbert Wiener (1948)



calculus ratiocinator

Self-reproducing automata

The General and Logical Theory of Automata

J. von Neumann

Von Neumann, J., 1951. In: Jeffress, L.A. (Ed.), *Cerebral Mechanisms of Behavior: The Hixon Symposium.* John Wiley and Sons, New York, pp. 1–41.

- **Living organisms can self-reproduce:** they produce structures of equal or even increased complexity during evolution.
- However, **Artificial machines degenerate in complexity during production** (factories are more complex than the machines they produce: production, control etc).

- **Is it possible to conceive a machine/automaton that self-replicates?**
- **What is the threshold of complexity above which machines can self-replicate?**

Self-reproducing automata

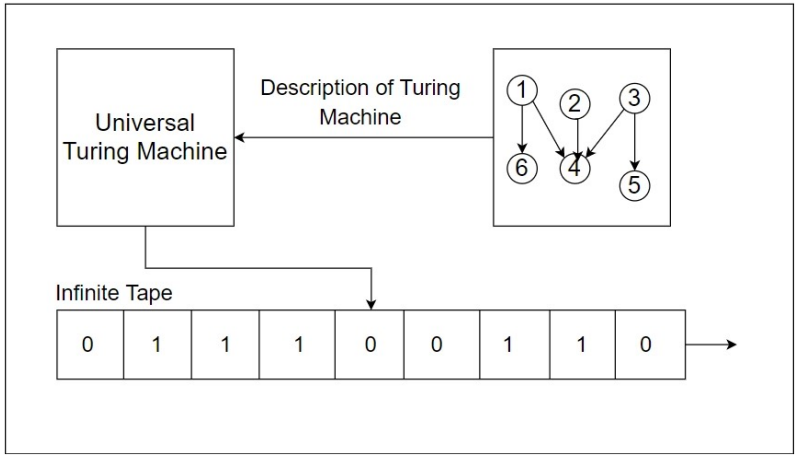
The General and Logical Theory of Automata

A. Turing (1912-1954)

Turing machine is an abstract entity:

- **Tape** (infinite)
- Reading/writing **Head**
- **State register**
- **Transition table** (a function of what is read and state)

Turing, A. M. *Proc. Lond. Math. Soc.* s2–42, 230–265 (1936).



Turing's theory of computing automata:

An automaton is able to "form" a certain sequence if it is possible to specify a finite length of tape, feed it to the automaton such that it will write the sequence on the tape.

The finite piece of tape constitutes the "instruction" of the automaton for this problem.

An automaton is "universal" if it simulates all possible Turing machines: any sequence that can be produced by any automaton (a Turing machine) can also be solved by this particular automaton that simulates it.

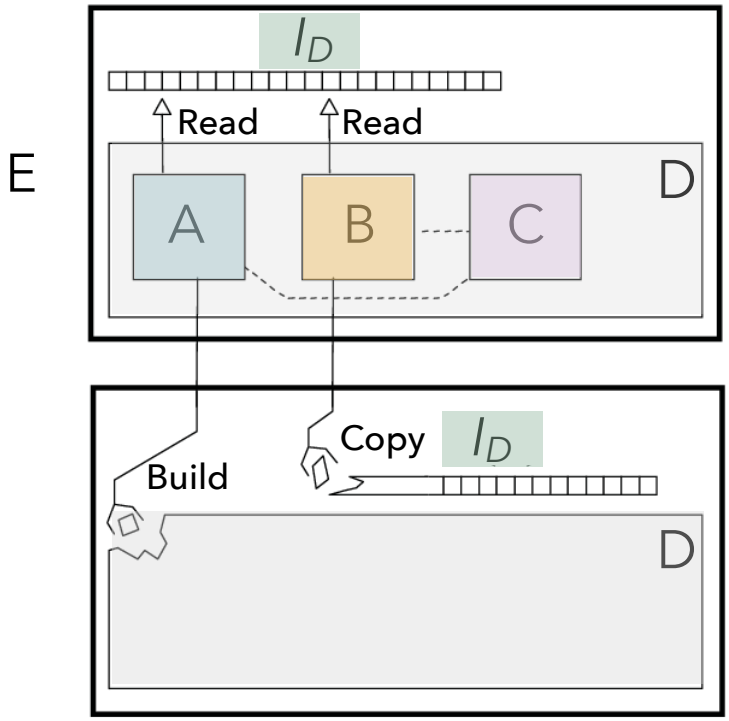
The Main Result of the Turing Theory.

We might expect a priori that this is impossible. ***How can there be an automaton which is at least as effective as any conceivable automaton, including, for example, one of twice its size and complexity?***

This automaton, which is constructed to read a description and to imitate the object described, is then the universal automaton in the sense of Turing.

Self-reproducing automata

- Requirements (to avoid degenerate complexity):
  - Copying the machine (A)
  - Copying the instructions to make the machine (B)



(a) **Automaton A, which when furnished the description of any other automaton** in terms of appropriate functions, **will construct that entity**.

A description in this sense will be called an instruction and denoted by a letter *I*

(b) **Automaton B, which can make a copy of any instruction *I***

that is furnished to it.

This automaton is nothing more subtle than a **« reproducer »**. (c)

(c) Combine the automata A and B with each other, and **with a control mechanism C.**

C will first cause A to construct the automaton which is described by this instruction *I*. Next C will cause B to copy the instruction *I*, and insert the copy into the automaton, which has just been constructed by A. Finally, C will separate this construction from the system A + B + C.

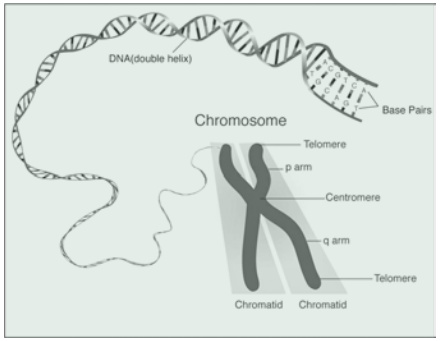
(d) denote D = A + B + C. D requires an instruction *I*.

Form an instruction I_D, which describes this automaton D, and insert to into A within D. Call the aggregate which now results E.

E is self-reproductive

$$E = D + I_D = A + I_D + B + C$$

Self-reproducing automata and cells

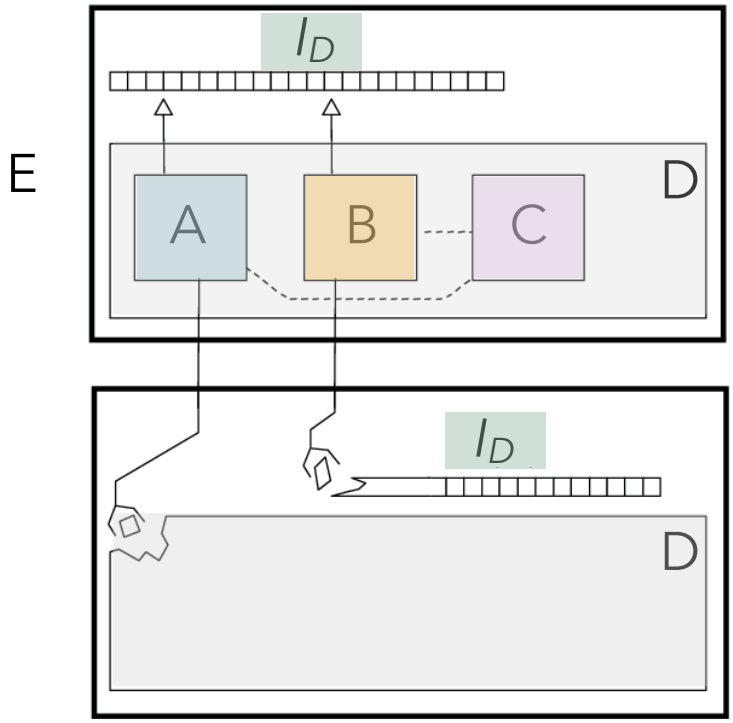


Interpretations of This Result and of Its Immediate Extensions. The description of this automaton *E* has some further attractive sides, into which I shall not go at this time at any length. For instance, it is quite clear that the instruction I_D is roughly effecting the functions of a gene. It is also clear that the copying mechanism B performs the fundamental act of reproduction, the duplication of the genetic material, which is clearly the fundamental operation in the multiplication of living cells. It is also easy to see how arbitrary alterations of the system *E*, and in particular of I_D, can exhibit certain typical traits which appear in connection with mutation, lethally as a rule, but with a possibility of continuing reproduction with a modification of traits.

Universal Self-Replicator (1948)	Biological Instantiation
I_D **Instructions** for automata	DNA (genes + …) (1953)
A **Automaton that constructs** an automaton based on instruction I_D	**Transcription (1961)** **Codon (1961), Translation machineries (mid 50s)** (DNA, mRNA, RNA polymerase, ribosomes)
B **Automaton that copies** I_D	**Replication machinery (1956)** DNA polymerase, topoisomerase, etc
C **Control module for A+B**	**Gene regulation (1961)** **Signal transduction (70-80s)**
D **Universal Turing Machine**	**Whole cellular machinary** excluding DNA
E **Self-Replicator** $E = D + I_D = A + I_D + B + C$	**Cell**

Self-reproducing automata and cells

Are the instructions/is the information strictly in the DNA?
Is the information complete in the genome?



- I_D encodes A, B and C.
- This picture suggests that A, as it builds D, the cell, provides building blocks that, with an energy source, self-assemble or self-organise into a cell (membrane, organelles etc).
- However, a cell does not strictly self-organise. It requires a structured template on top of chemical components.
- Thought experiment: grind a cell into chemical condensate. Does it reform?
- Structural heredity provides additional source of information at any time (see course #5 -10 dec)
- These structures result from evolution of cellular mechanisms

The genome as a blueprint/code-script ?

- **What in the genome is a de facto a blueprint:**

The coding sequence of proteins

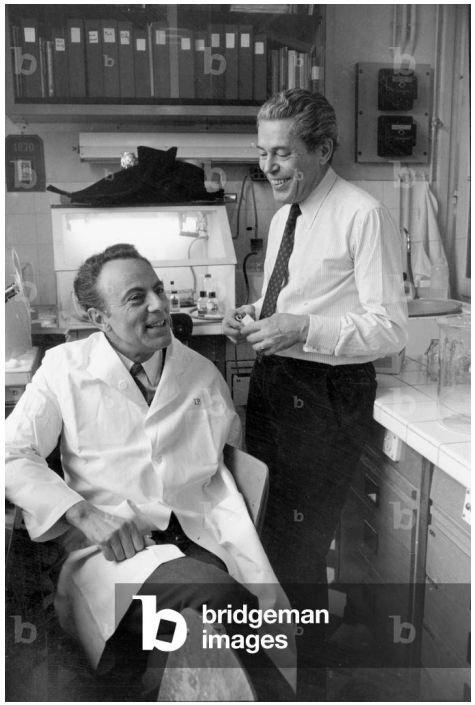
There is a direct, linear mapping of DNA into RNA and Proteins.

- **What is *not* a blueprint:**

The <u>means of controlling the execution of the code</u> in space and time.

This requires a whole cellular and organismal environment.

The genome as a *program*



REVIEW ARTICLE

Genetic Regulatory Mechanisms in the Synthesis of Proteins †

FRANÇOIS JACOB AND JACQUES MONOD

Services de Génétique Microbienne et de Biochimie Cellulaire, Institut Pasteur, Paris

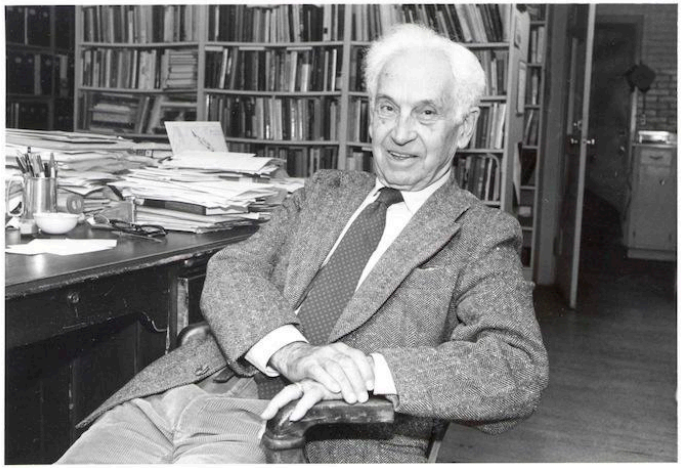
J. Mol. Biol. (1961) **3**, 318–356

Cause and Effect in Biology

Kinds of causes, predictability, and teleology are viewed by a practicing biologist.

Ernst Mayr

Science (1961), 134: 1501-1506



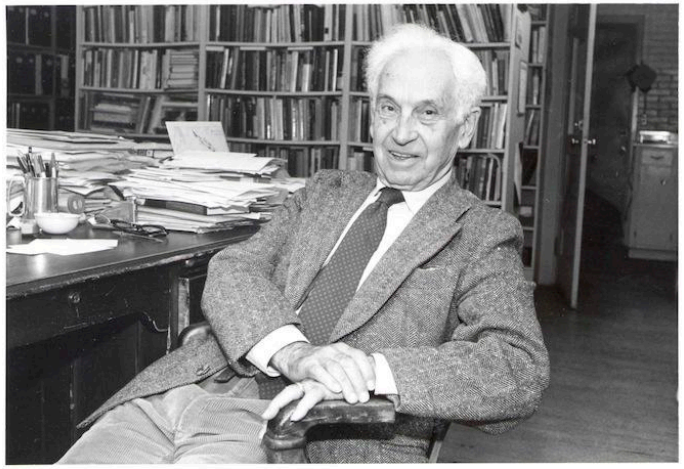
The genome as a *program*

Cause and Effect in Biology

Kinds of causes, predictability, and teleology are viewed by a practicing biologist.

Ernst Mayr

Science (1961), 134: 1501-1506



- Functional biology: proximal causes. How? « operation and interaction of structural elements, from molecules to organs and whole organism ».
- Evolutionary biology: ultimate causes. Why, ie. How come? Teleonomy (internal purposiveness).
- Suggests a sequence of coded software instructions.

We can use the language of information theory to attempt still another characterization of these two fields of biology. The functional biologist deals with all aspects of the decoding of the programmed information contained in the DNA code of the fertilized zygote. The evolutionary biologist, on the other hand, is interested in the history of these codes of information and in the laws that control the changes of these codes from generation to generation.

4) The existence of complex codes of information in the DNA of the germ plasm permits teleonomic purposiveness.

But let us not have an erroneous concept of these codes. It is characteristic of these genetic codes that the programming is only in part rigid. Such phenomena as learning, memory, nongenetic structural modification, and regeneration show how "open" these programs are. Yet, even here there is great specificity, for instance with respect to what can be "learned," at what stage in the life cycle "learning" takes place, and how long a memory engram is retained. The program, then, may be in part quite unspecific, and yet the range of possible variation is itself included in the specifications of the code. The codes, therefore, are in some respects highly specific; in other respects they merely specify "reaction norms" or general capacities and potentialities.

The genome as a program - *Lac* operon

- Jacob & Monod:
- The *Lac* operon and phage λ system as paradigm of regulation of protein synthesis

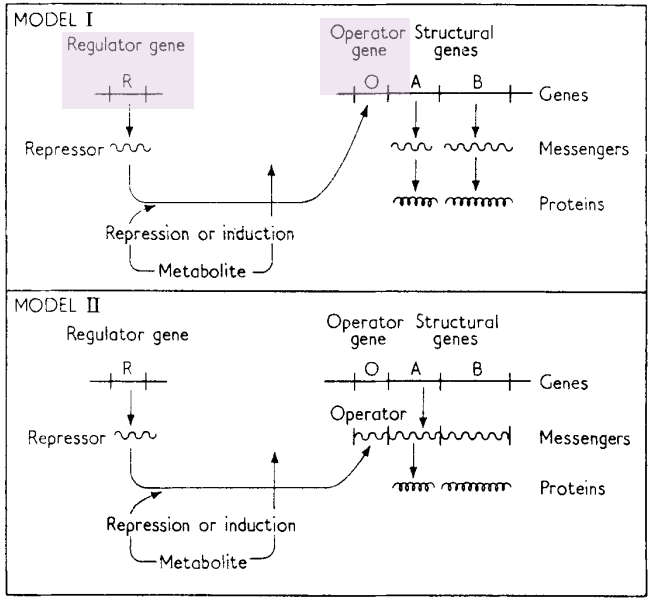


FIG. 6. Models of the regulation of protein synthesis.

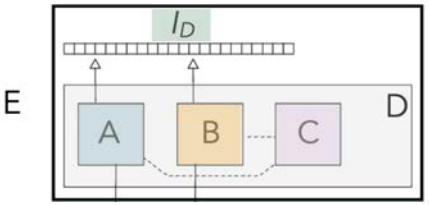
J. Mol. Biol. (1961), 3, 318-356

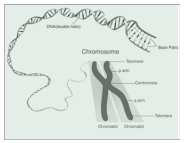
(a) That the mechanisms of control in all these systems are negative, in the sense that they operate by inhibition rather than activation of protein synthesis.

(b) That in addition to the classical structural genes, these systems involve two other types of genetic determinants (regulator and operator) fulfilling specific functions in the control mechanisms.

(c) That the control mechanisms operate at the genetic level, i.e. by regulating the activity of structural genes.

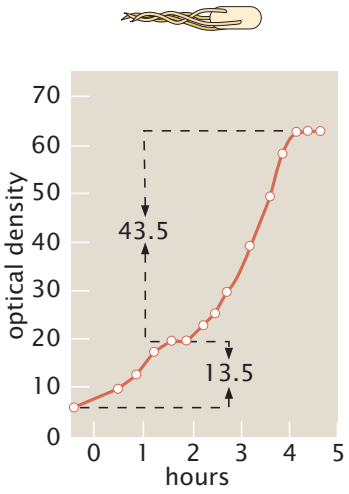
- Discovery of genes that regulate other genes (Regulator)
- Discovery of sequences that respond to Repressor (Operator)
- The control module C of von Neumann's self-replicator includes proteins (ie. Repressor), and non-coding regulatory sequences (Operator)

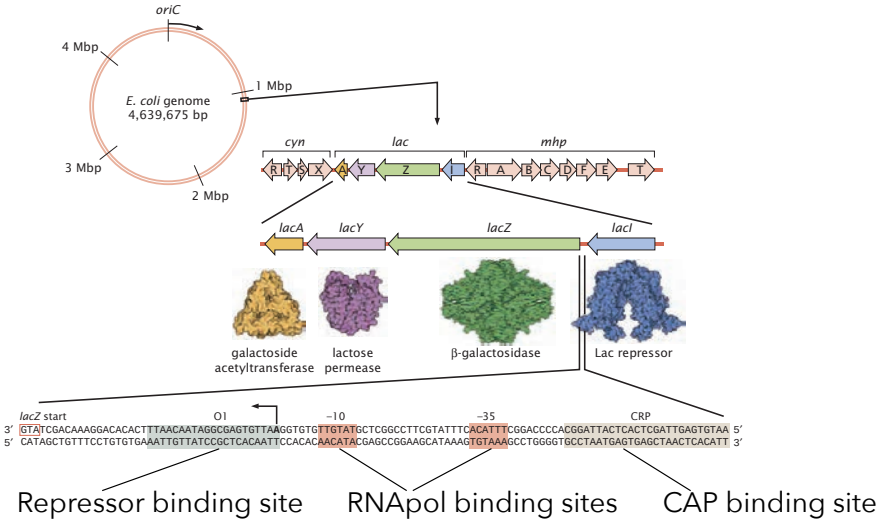




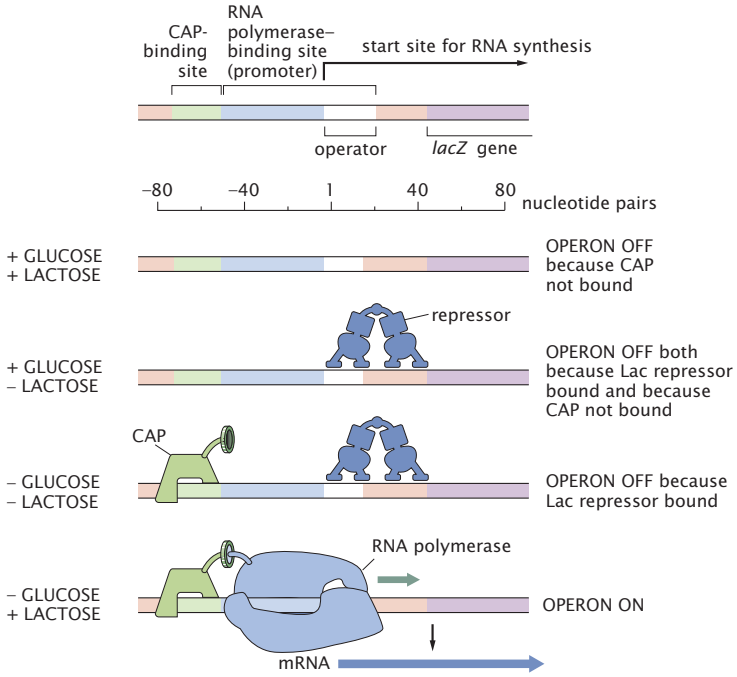
The genome as a program - *Lac* operon

- *E. coli* growth curve: 2 phases of growth based on glucose first and then lactose.
- This entails induction of genes that encode for proteins that 1) digest lactose (ß-Gal), 2) induce transport across the membrane (permease) etc.





- **In presence of lactose, the Lac repressor is repressed (does not bind the operator).**
- In presence of glucose, cAMP is low and CAP cannot bind DNA



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

Universality of gene regulation

- Repression and activation of many genes
  – molecular cell differentiation, spatial cell differentiation (pattern formation) during embryonic development

The fundamental problem of chemical physiology and of embryology is to understand why tissue cells do not all express, all the time, all the potentialities inherent in their genome. The survival of the organism requires that many, and, in some tissues most, of these potentialities be unexpressed, that is to say *repressed*. Malignancy is adequately described as a breakdown of one or several growth controlling systems, and the genetic origin of this breakdown can hardly be doubted.

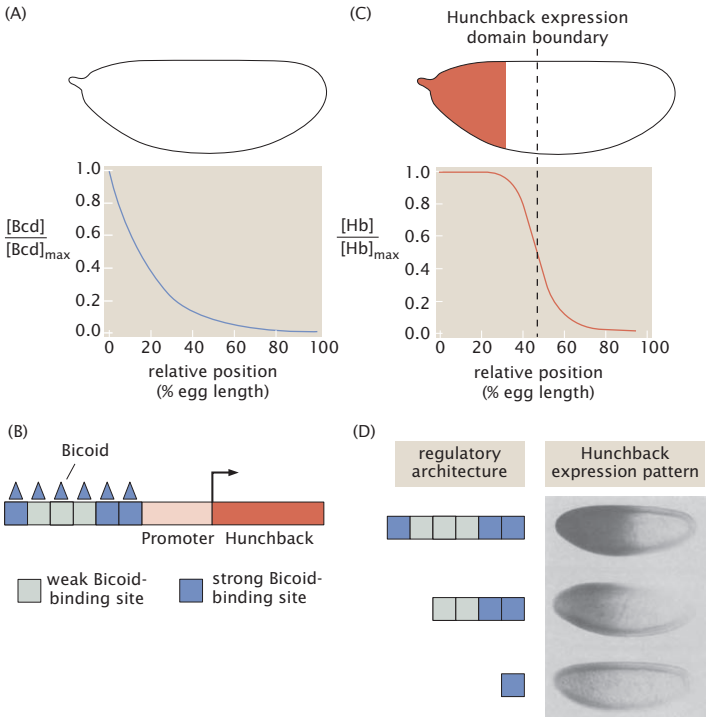
According to the strictly structural concept, the genome is considered as a mosaic of independent molecular blue-prints for the building of individual cellular constituents. In the execution of these plans, however, co-ordination is evidently of absolute survival value. The discovery of regulator and operator genes, and of repressive regulation of the activity of structural genes, reveals that the genome contains not only a series of blue-prints, but a co-ordinated program of protein synthesis and the means of controlling its execution.

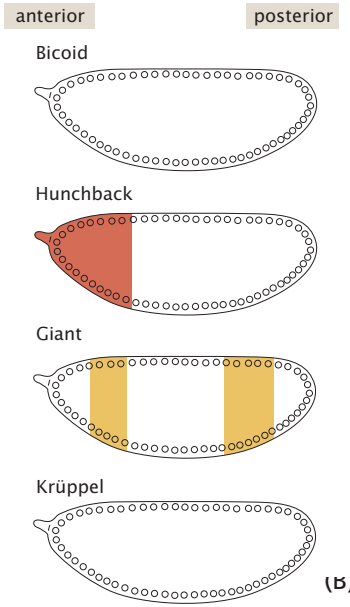
- Suggests a sequence of coded software instructions
- Or of a preset agenda

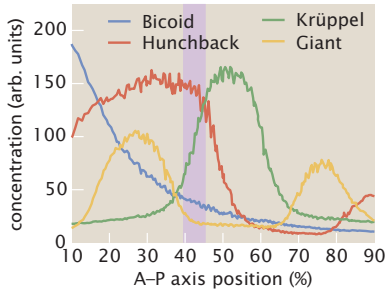
F. Jacob & J. Monod. *J.Mol.Biol.* 3: 318-356 (1961)

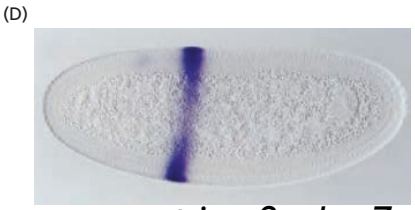
The genome as a program - developmental spatial patterning

- Spatial patterning in the early *Drosophila* embryo:
- The *responsiveness* to spatial molecular cues (regulators) is encoded in regulatory sequences
- This leads to spatial regulation of gene expression (patterning)

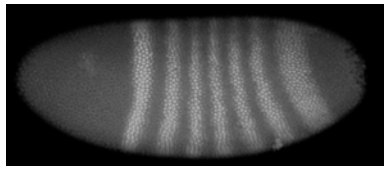




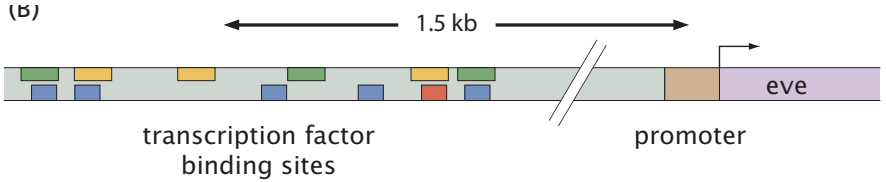




eve stripe 2 – LacZ



Even-skipped (Eve)



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

The genome as a program - developmental spatial patterning

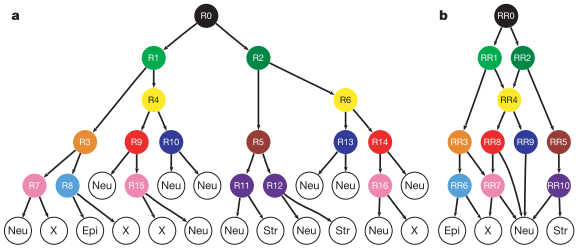
- Development as a genetic and cellular automaton controlling cell identity
- Garcia Bellido: distinction between *selector* genes, and *cytodifferentiation* genes

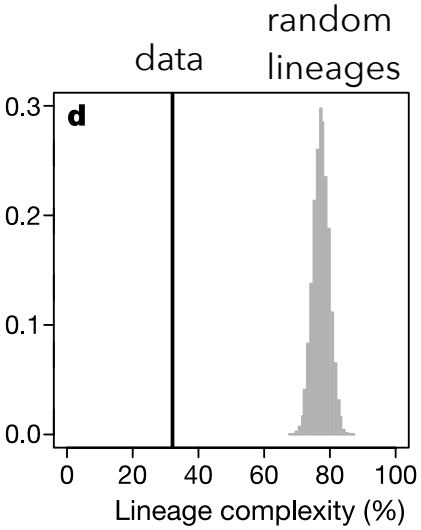
« Genes of a first group (cyto-differentiation genes) would include those controlling cell behaviour relevant to morphogenesis and common to most developing systems: mitotic rate, mitotic orientation, cell recognition and cuticular differentiation.
Those of a second group (**selector genes**) seem to control developmental pathways and share several operational characteristics. A functional scheme is advanced showing how selector genes may become activated and control development. We postulate that inductor molecules interfere with the products of activator genes which are selector specific. In this way signals extrinsic to the genome become translated into genetic ones. The activation, or repression, of selector genes occurs once in development and remains clonally irreversible ».

García-Bellido A. *Ciba Found Symp*. 1975;0(29):161-82.

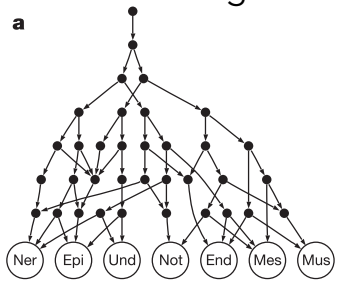
The genome as a program - developmental spatial patterning

- Development as a genetic and cellular automaton controlling cell identity
- **Cell lineage algorithmic complexity:** ratio between minimal reduced rules to generate the lineage (11) and the actual number of cell divisions (17): 65%.
- Embryonic lineage complexity near that of in silico evolved lineages constrained by spatial positioning of cells



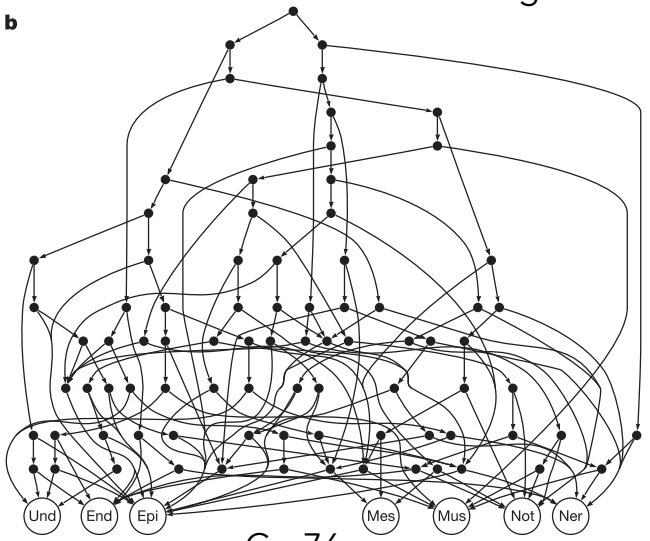


H. roretzi lineage



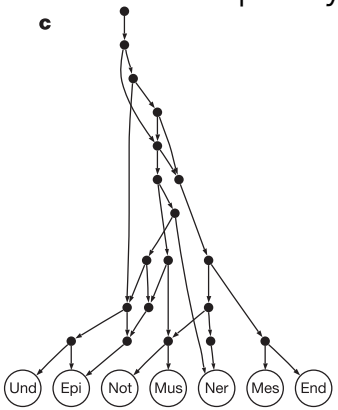
C= 32

A random bifurcation lineage



C= 76

Minimal complexity

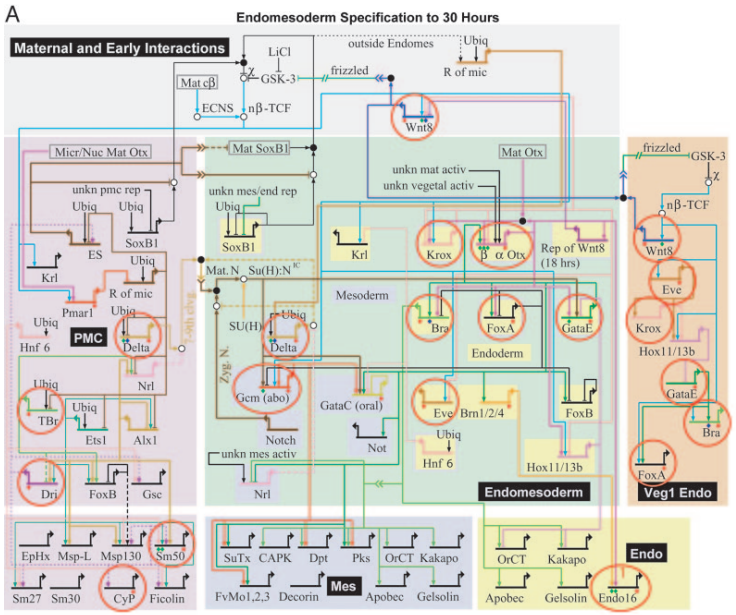


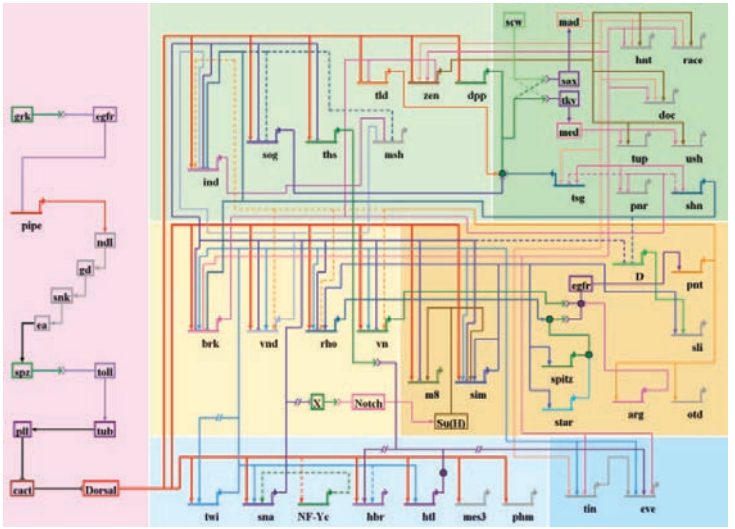
C= 17

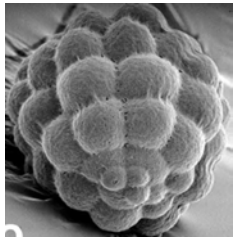
See also: JA. Alsous, P. Villoutreix et al. J. Dunkel, Nature Physics 2018
doi.org/10.1038/s41567-018-0202-0

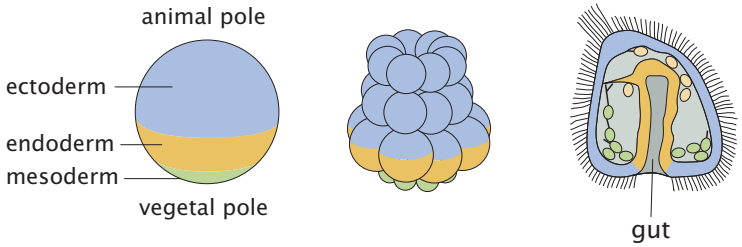
Azevedo RBR et al. A. Leroi, *Nature*, 433: 152-156 (2005)

The genome as a program - gene regulatory networks

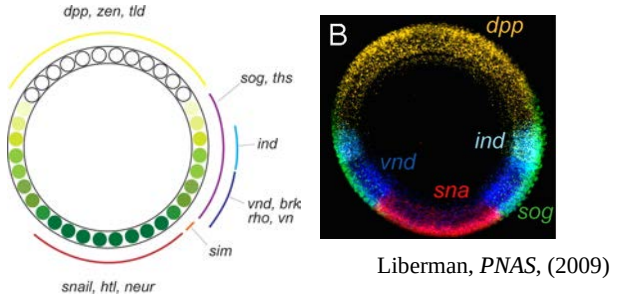








L. Bodenstein. *Mechanisms of Development*, 162 (2020)
https://doi.org/10.1016/j.mod.2020.103606



Liberman, *PNAS*, (2009)

Critiques and Limits of these « genome encoding » metaphors

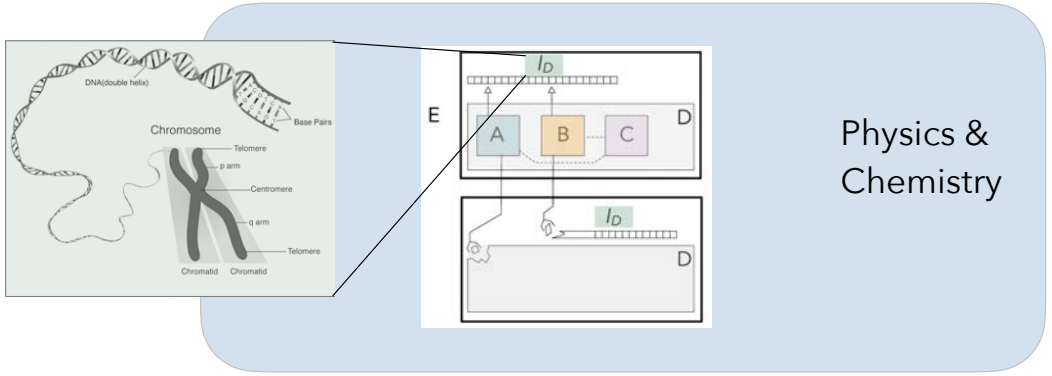
- **Blueprint:** The genome is not isomorphic to the outcome
  Does not specify « how to build » the organism

- Codescript
- Program
  - **Algorithmic determinism is implied but organisms can find alternative routes when perturbed (eg. regeneration, adaptation)**
  - **A computer program is brittle** (high probability of failure if error in code) but **organism are highly robust** and can repair errors.
  - **Organisms are subject to fluctuations** of all kinds unlike machines/ computers piloted by programs
  - **Organisms reconfigure constantly** unlike most machines/programs.
  - **The organism/cellular states are not unambiguously defined in a complete way** neither by genes nor by anything else.

> The natural gene does probably not contain a complete description of the object whose construction its presence stimulates. It probably contains only general pointers, general cues.
>
> **J. von Neumann**

>>Need to characterise the properties of these « pointers, general cues ».

The genome as a recipe



- Description in the genome of the core components and of the means to synthesize them in the right sequence.
- But the actual making of cells and organisms is set and constrained by the physical-chemistry
- **The physics and chemistry is not encoded but provides the necessary environment for the genome and the whole cell to function at a molecular level and across scales**
- Conveys a less deterministic concept than the algorithmic program and states what the genome doesn't encode (eg. Physics and chemistry).
- Yet, it doesn't present what the properties of the « recipe » are that would distinguish it from the program

K.J. Mitchell (2018). *Innate — How the wiring of our brains shapes who we are.* Princeton Univ. Press, Princeton NJ.

Plan

1. The egg as a « compressed information » state
2. The genome as a carrier of developmental information
3. Metaphors for the genome: blueprint, code script, program, etc.
4. Properties of Genotype to Phenotype mapping
5. Low dimensional representations

Relationships between Genotype and Phenotype

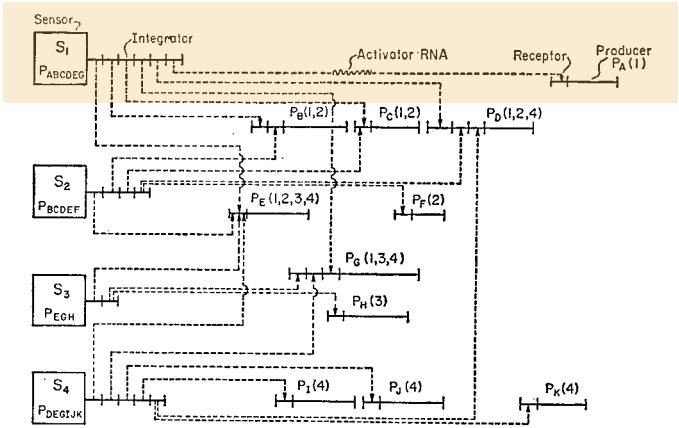
- One gene one function hypothesis
  – Ephrussi and Beadle: *Drosophila* eye colour phenotypes and mutants suggested 1 gene - 1 enzyme relationship
  – Beadle and Tatum (amino acids synthesis mutant in *Neurospora crassa*): 1 gene - 1 enzyme. Nobel Prize, 1958.

B. Ephrussi and G. Beadle (1935-36)

- Multiple genes contribute to a phenotype (function)
- A function (phenotype) is dependent on many genes organised in Regulatory Network (GRNs)

Table 1. Several of the functionally linked enzyme systems present in liver (*17*, chapter 12; *36*). Uridine monophosphate, UMP; adenosine monophosphate, AMP.

System	Number of enzymes
Glycogen synthesis	5
Galactose synthesis	6
Phosphogluconate oxidation	11
Glycolysis	12
Citric acid cycle	17
Lecithin synthesis	8
Fatty acid breakdown	5
Lanosterol synthesis	10
Phenylalanine oxidation	8
Methionine to cysteine	10
Methionine to aspartic acid	10
Urea formation	10
Coenzyme A synthesis	6
Heme synthesis	9
Pyrimidine synthesis (to UMP)	6
Purine synthesis (to AMP)	14



*GENETIC CONTROL OF BIOCHEMICAL REACTIONS IN NEUROSPORA**

BY G. W. BEADLE AND E. L. TATUM
BIOLOGICAL DEPARTMENT, STANFORD UNIVERSITY

Communicated October 8, 1941

Summary

A theory for the genomic regulation systems of higher organisms is described. Batteries of producer genes are regulated by activator RNA molecules synthesized on integrator genes. The effect of the integrator genes is to induce transcription of many producer genes in response to a single molecular event.

Eric Davidson, *Science.* 165: 349-357 (1969)
Gene regulation for higher cells: A theory

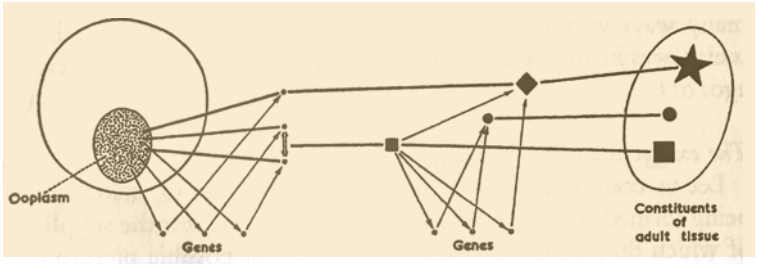
Relationships between Genotype and Phenotype

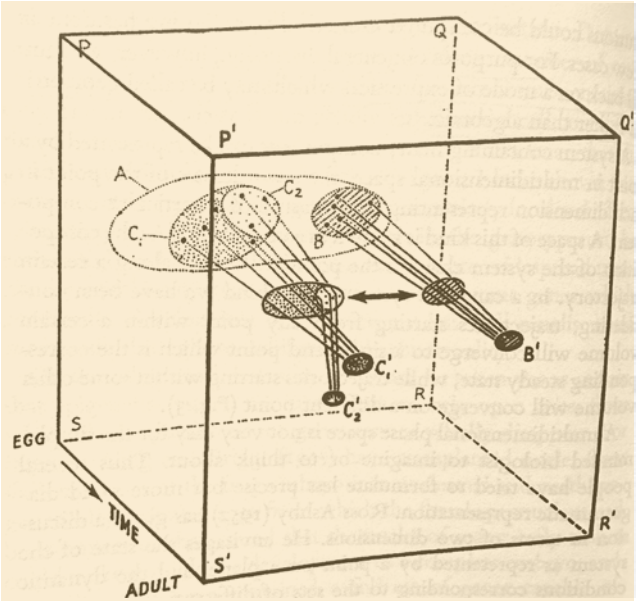
- The Epigenetic landscape concept

The strategy of the genes (1957). Chap. 2. *The cybernetics of development*

- Development is complex process comprising regionalisation (patterning), histogenesis (differentiation) and morphogenesis.
- Development entails **evolution overtime in a multidimensional space** that characterises its composition (genes, proteins and other components of cytoplasm).
- **A phase space best characterises development.**

C.H. Waddington
(1905-1975)





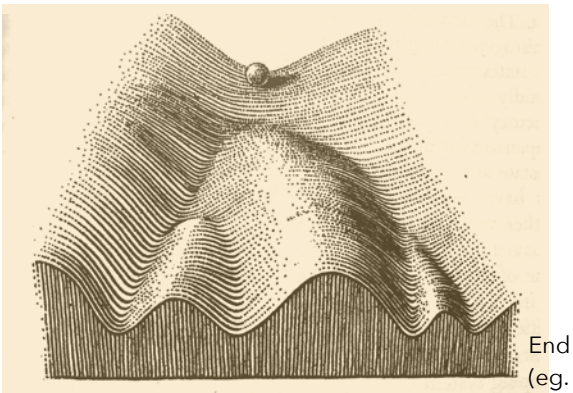
C.H. Waddington (1957). *The strategy of the genes. A discussion of some aspects of theoretical biology.*

Relationships between Genotype and Phenotype

- Epigenetic landscape concept

- This system exhibits tendency towards a kind of **equilibrium centred not on state but on a direction of change** (homeorhesis, flow)
- **A creode (« necessary path ») is a representation by a trajectory in phase space of a temporal succession of states towards which the system will relax if perturbed.**

Initial states (different cytoplasmic states in different parts of the egg)



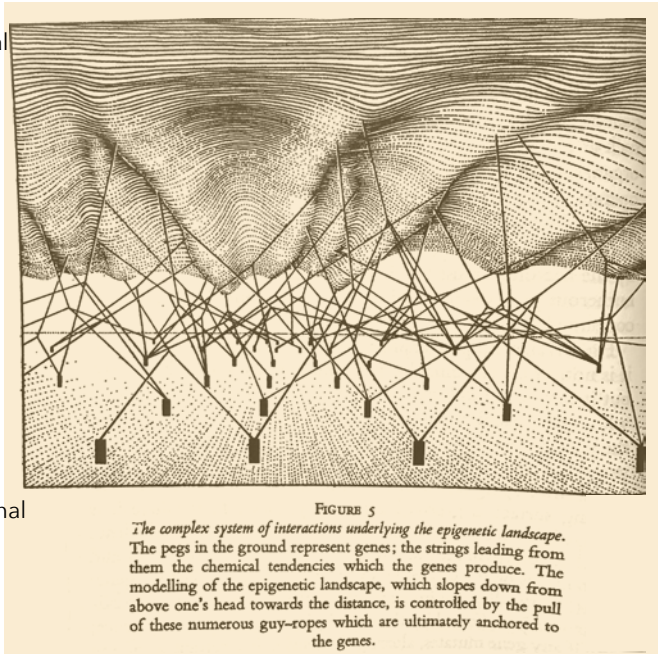
End states (eg. Eye, brain, spinal chord …)

FIGURE 4

Part of an Epigenetic Landscape. The path followed by the ball, as it rolls down towards the spectator, corresponds to the developmental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

- A complex system of interactions underlies the epigenetic landscape
- Pegs represent genes and tension on guy ropes the chemical forces exerted by genes
- Connection between genotype and phenotype is Non-isomorphic, non-linear, combinatorial, indirect

Phenotype
Low dimensional



Genotype
High dimensional

FIGURE 5

The complex system of interactions underlying the epigenetic landscape. The pegs in the ground represent genes; the strings leading from them the chemical tendencies which the genes produce. The modelling of the epigenetic landscape, which slopes down from above one's head towards the distance, is controlled by the pull of these numerous guy-ropes which are ultimately anchored to the genes.

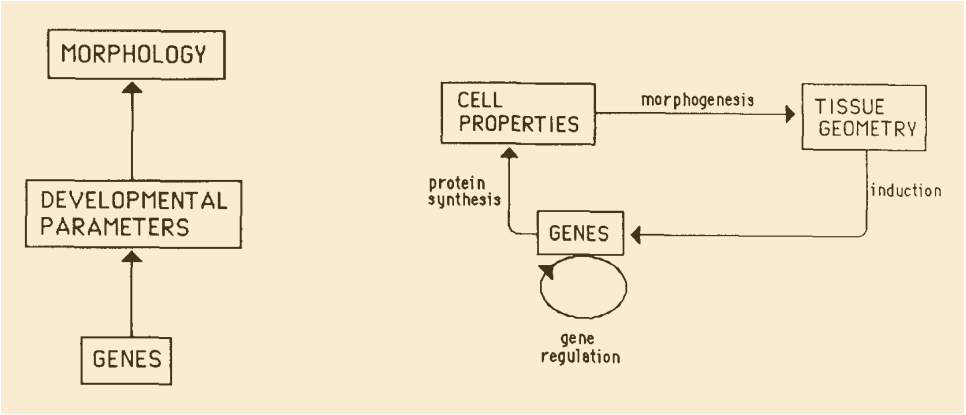
The Genotype to Phenotype *mapping*

Pere Alberch
(1954-1998)

From genes to phenotype: dynamical systems and evolvability

P. Alberch
Museo Nacional de Ciencias Naturales (CSIC), José Gutiérrez Abascal, 2, 28006 Madrid, Spain

- Abandon the hierarchical view whereby genes control a sequence of processes that lead to form. If so development would be entirely dependent on genes. This is incorrect
- Instead consider feedback interactions between genes, cellular properties, tissue geometry, implicating complex physico-chemical processes.



Genetica **84**: 5–11, 1991.

The Genotype to Phenotype mapping

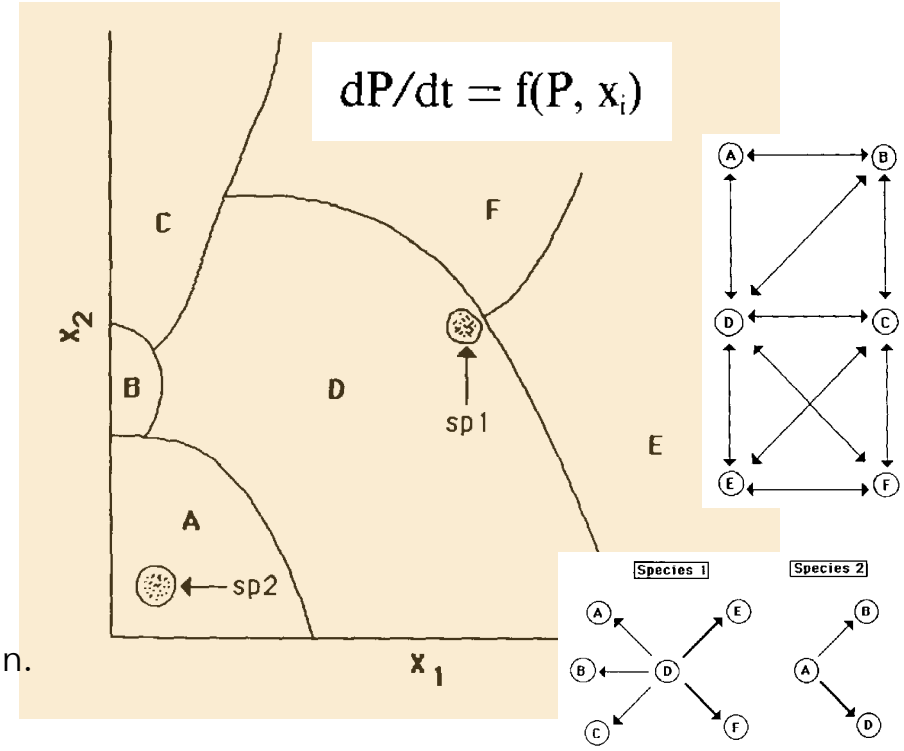
- <u>Hypothesis:</u>

1. **Many combinations of parameter values result in the same phenotype**, that is, there is no one-to-one correlation between genotype and phenotype.

2. The **stability of a phenotype is related to the area of its domain** in parameter space (canalisation).

3. The lines correspond to critical (x_1, x_2) values. They constitute **transformational boundaries** among phenotypes. **Bifurcation boundaries in dynamical system theory.**

4. The stability of a particular set of phenotypes will depend on its position in parameter space (sp1 and sp2).



Implications:

– **Robustness:** many genotypes give rise to same phenotype

– many genotypic changes can give rise to phenotypic transformation.

– **Evolvability:** Transformational diagram define allowed phenotypic changes, with probability set by the length of boundaries in phase space.

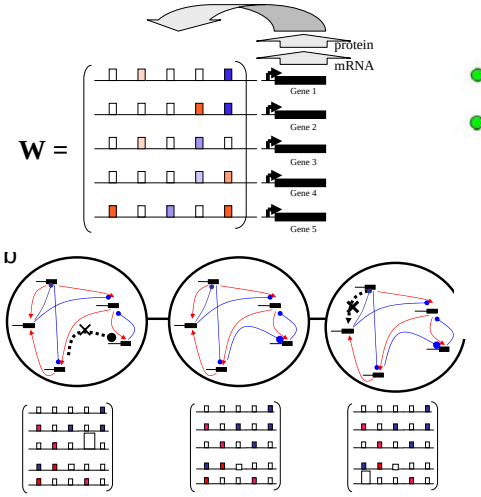
Genetica **84**: 5–11, 1991.

The Genotype to Phenotype mapping

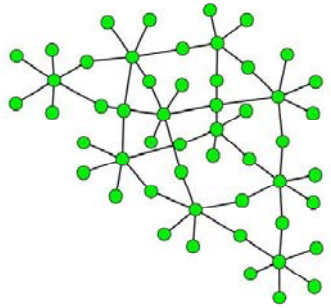
- Computational study:

Innovation and robustness in complex regulatory gene networks

S. Ciliberti*, O. C. Martin*†, and A. Wagner‡§



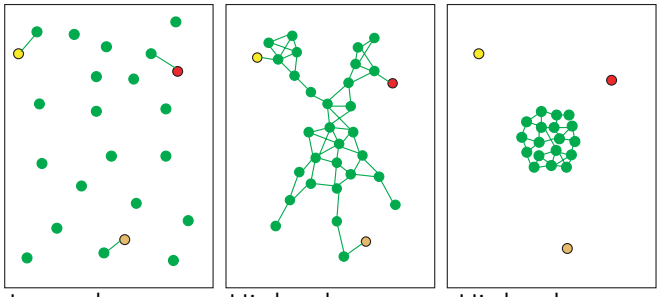
Nodes with given topology



Neutral network of given phenotype

(nodes differ by 1 regulatory interaction)

- The phenotype is the gene expression pattern produced by regulatory interactions in a network
- Networks with very different organizations can have the same phenotype: neutral network .
- In contrast, two networks with completely unrelated phenotypes can be found very close to each other in genotype space.
- The genotype space can be traversed in small steps without changing the phenotype: scattered but connected genotypes
- Different phenotypes become accessible in different parts of genotype space.
- This is crucial for evolutionary innovation in gene expression patterns.



Low robustness Low innovation | High robustness High innovation | High robustness Low innovation

Ciliberti,S., Martin,O.C. & Wagner, A. *PNAS* 104, 13591–13596 (2007).

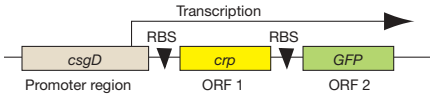
The Genotype to Phenotype mapping

- <u>Experimental studies:</u> rewiring bacterial gene networks.

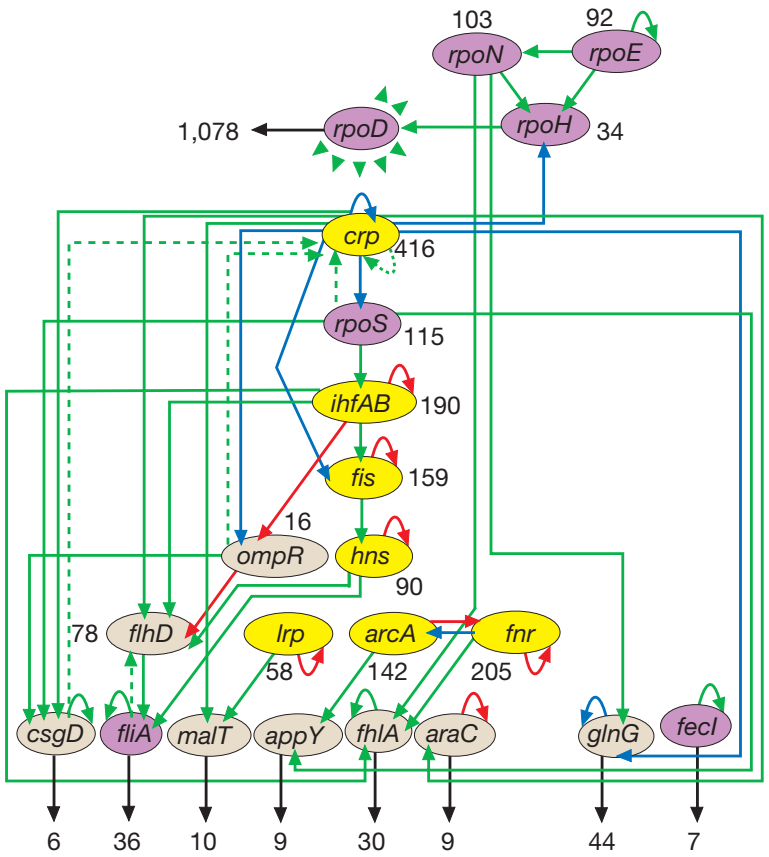
Evolvability and hierarchy in rewired bacterial gene networks

Mark Isalan[1], Caroline Lemerle[2], Konstantinos Michalodimitrakis[1], Carsten Horn[2], Pedro Beltrao[2], Emanuele Raineri[1], Mireia Garriga-Canut[1] & Luis Serrano[1]

- 598 new regulatory connections between regulatory promoters and sequence of transcription factors



- 95% of such new network « mutations » do not show any phenotype: robustness



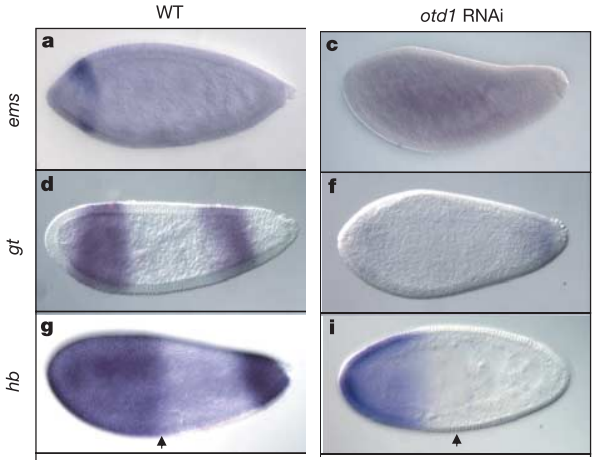
Isalan,M. et al. and L. Serrano. *Nature* 452, 840–845 (2008).

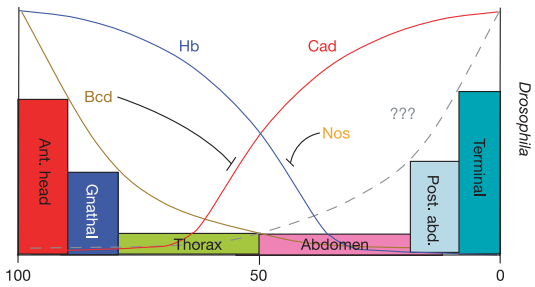
The Genotype to Phenotype mapping

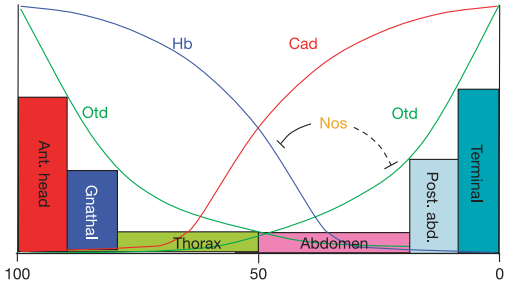
- <u>Experimental studies:</u> Innovations in GRNs controlling embryonic patterning

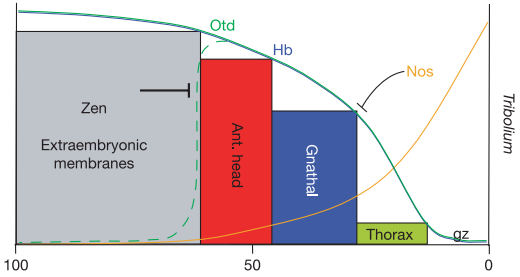
- Studies of embryonic patterning in insects revealed evolutionary changes in GRNs
- In *Drosophila* the morphogen Bcd controls spatial expression of gap genes *gt and hb*
- In other insects, eg. The wasp *Nasonia,* and the beetle *Tribolium* Bcd is absent and this function is performed by maternally deposited Otd.

Otd protein







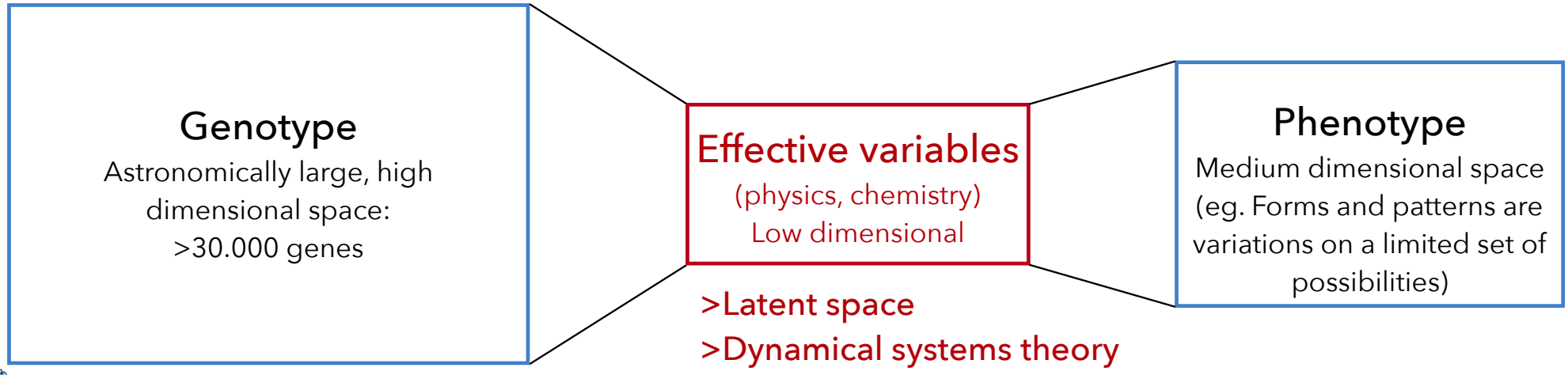


Lynch, J.A., et al. & Desplan C. *Nature* 439, 728–732 (2006).
Schröder, R. *Nature* 422, 621–625 (2003).

Summary

- The code-script/program/recipe metaphors all fail to capture what is the format of the information or representation of the organism in the genome of the egg.

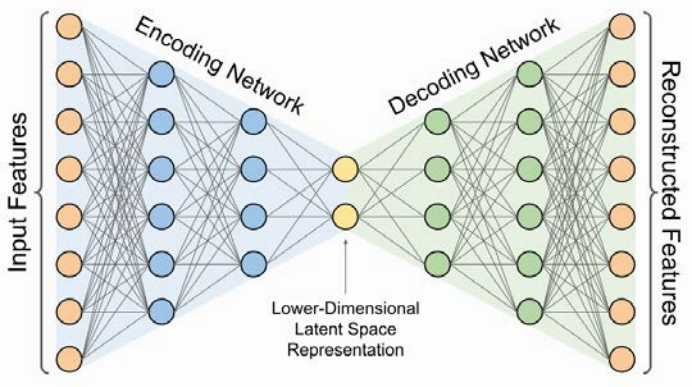
- The non-linear, non isomorphic and indirect genotype-phenotype mapping argues that **the genome contains a representation that:**
  – characterises and constrains the dynamics of a complex system
  – defines the effective parameters that encode such dynamics
  – is a low dimensional space



Another computational metaphor: The genome as a generative model?

- Neural networks and Autoencoders: developed in 90s for dimensionality reduction
- **Constructs (encodes) a low dimensional representation of input data: Latent space.**
  Latent variables capture the statistical regularities in input data.
- **It constructs together a generative model that is used to Decode the latent space.**

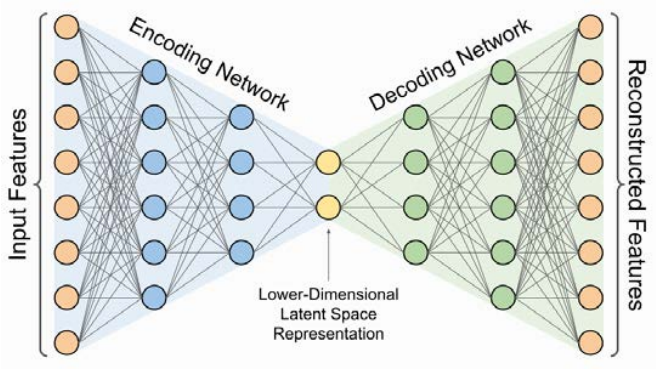
- **Autoencoders encode latent variables and the means to decode them.**
- Shows some important features of genotype to phenotype mapping (non-linear, non-isomorphic/ distributed, indirect)
- Variational auto encoders link probability distributions of variables, thereby better preserve distance and allow generalisation in representation.

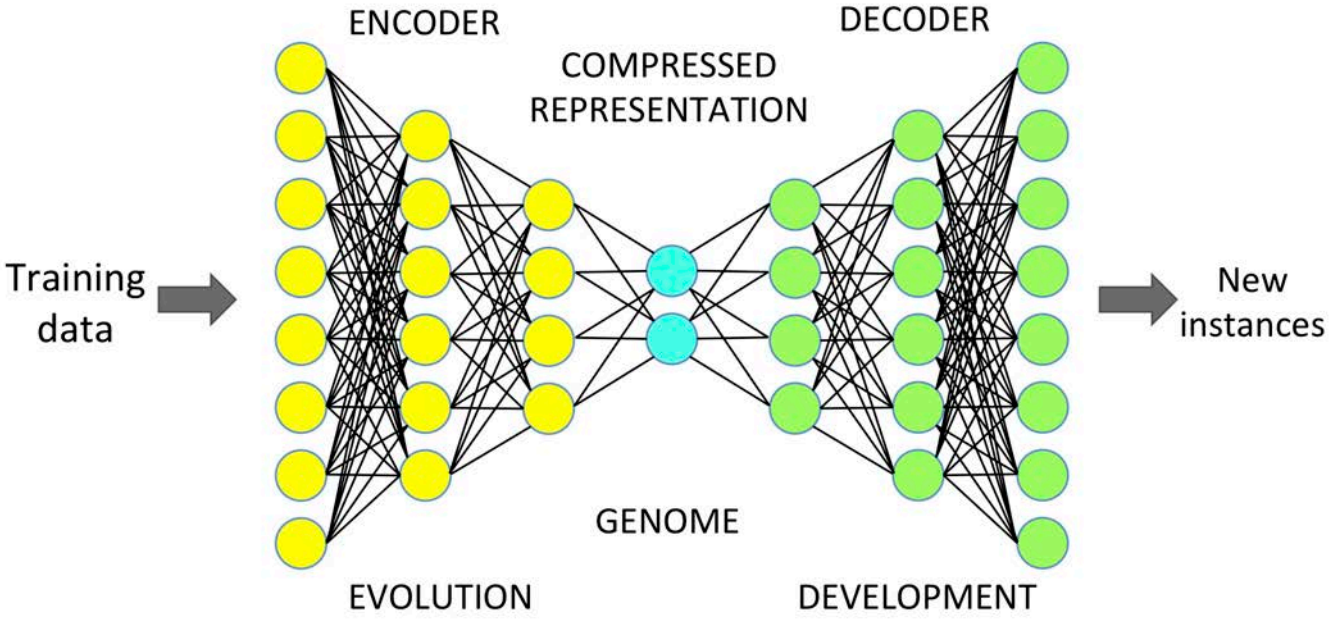


https://www.assemblyai.com/

Another computational metaphor: The genome as a generative model?

- Autoencoders encode latent variables and the means to decode them.
- Shows some important features of genotype to phenotype mapping (non-linear, non-isomorphic/distributed, indirect)
- Evolution is the Encoder
- Development is the Decoder

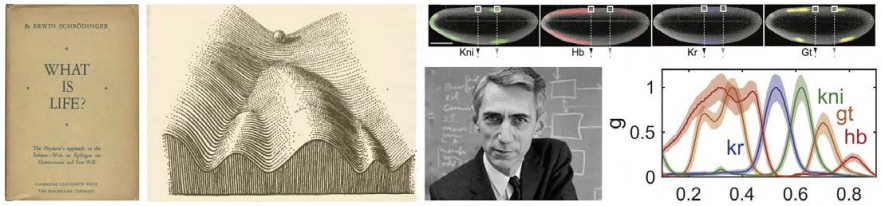




K. Mitchell and N. Cheney
arXiv:2407.15908v1

Summary

- The non-linear, non isomorphic and indirect genotype-phenotype mapping argues that **the genome contains a representation that:**
  – **characterises and constrains the dynamics** at different scales.
  – **defines the effective parameters that encode such dynamics**
  – is a low dimensional space
- The encoding of dynamics and of steady states in the genome is **decoded in the physical environment of cells and cell ensembles during development: genetics, mechanics and geometry.**
  (courses 2, 3, 4, 5).
  Thus the symbolic and physical features of living systems are intertwined (unlike algorithmic machines).
- **This involves feedback interactions, learning processes and memory of past states/trajectories (course 6)**



Thomas LECUIT, chaire Dynamiques du vivant

Qu'est-ce que l'information biologique ?

COURS : 12 novembre > 17 décembre 2024

COURS

De 10h à 11h30
Amphithéâtre Guillaume Budé

Mardi 12 novembre 2024
Introduction :
quelles représentations pour le génome ?

Mardi 19 novembre 2024
Codes biologiques

Mardi 26 novembre 2024
Encodage, décodage
et représentations de *l'espace*

Mardi 3 décembre 2024
Encodage, décodage
et représentations du *temps*

Mardi 10 décembre 2024
Information structurelle et géométrique

Mardi 17 décembre 2024
Mémoires et apprentissages

COLLOQUE

De 9h à 18h
Amphithéâtre Maurice Halbwachs

Vendredi 16 mai 2025
Information Processing in Biological Systems

Les cours et colloques
sont gratuits, en accès libre,
sans inscription préalable.