Complex Systems Summer School

Introduction to complex systems

Guillaume Beslon

INSA – LIRIS – IXXI

Introduction

- Aim of the lecture:
  - Introduction to complex systems and complex systems science
  - Present my own work in “digital genetics”
- Who am I?
  - Guillaume BESLON (guillaume.beslon@inria.fr)
  - Professor at the INSA-Lyon, LIRIS Lab. (Laboratoire d’Informatique en Image et Systèmes d’Information)
  - Head of the INRIA « Beagle » Team (Computational Biology and Artificial Evolution)
  - Co-director of IXXI (Rhône-Alpes Complex Systems Institute)
  - Research topics: Individual-based modeling of complex biological systems (mainly evolution), artificial life

Schedule (tentative)

- Part 1: What is a complex system?
  - Definition(s)
  - Why studying complex systems?
  - What is “complex systems science”?
    - Is it really different from “traditional” disciplinary science?
  - Is there a methodology in complex systems science?
- Part 2: Introduction to digital genetics
  - Modeling evolution to understand biological complexity
  - Illustration of the “methodology”
  - My own work …

Complex Systems Summer School

Part 1: What is a complex system?

Guillaume Beslon

INSA – LIRIS – IXXI

What is a complex system?

- Definitions are important because:
  - The term is widely used in science: "*I think the next century will be the century of complexity*" [S. Hawking, 2001]
  - The relationship of complex systems science with other sciences is often difficult (need to identify the differences)
  - Politics need maps to define global scientific policy
  - Complex systems are ubiquitous; Is there a global definition? A global question?
    - If no, there is no such thing as a complex systems science!
    - If yes, what is it?
- But, in fact, there is no universally accepted definition!
  - And lots of problems and conflicts!

Definition

- The latin root: “complexus”
  - Entangled, entwined, embracing …

The New Oxford American Dictionary

SECOND EDITION

com•plex

adjective |käm'pleks; kəm'pleks; 'käm,pleks|

1 consisting of many different and connected parts : *a complex network of water channels.*
- not easy to analyze or understand; complicated or intricate : *a complex personality* | *the situation is more complex than it appears.*

2 Mathematics denoting or involving numbers or quantities containing both a real and an imaginary part.

3 Chemistry denoting an ion or molecule in which one or more groups are linked to a metal atom by coordinate bonds.

[…]

ORIGIN mid 17th cent. (in the sense [group of related elements]): from Latin ***complexus***, past participle (used as a noun) of ***complectere 'embrace, comprise,'*** later associated with ***complexus 'plaited'***; the adjective is partly via French ***complexe***.

Definition

- Most definitions follow from these two basic properties

"A set of items interacting via simple local rules in which emergent properties cannot be directly deduced from the local rules"

[M. Morvan, founder of the IXXI]

"Complex systems are systems with multiple interacting components whose behavior cannot be simply inferred from the behavior of the components"

[NECSI]

"A system is complex if it exhibits nontrivial emergent and self-organizing behavior"

[M. Mitchell, SFI, 2009]

But

- All these definitions contain terms like:
  - “directly deduced”, “simply inferred”, “difficult to”, “essential property”, “emergent behavior”, “non-trivial behavior”, …
- All these terms either
  - Depend on the cognitive abilities of the observer and on his scientific knowledge
  - Are not really defined (or defined as antonyms of “complex”)
- These definitions are subjective, self-referent and, actually, dangerous!

“A complex system is a system in which large networks of components with no central control and simple rules of operation give rise to a complex collective behavior, sophisticated information processing, and adaptation via learning or evolution.”

[M.Mitchell, 2009]

So, what is a complex system?

(my definition)

- General agreement on:
  - The structure of the system (“many elements”)
  - Some subjective judgment (not always clearly accepted)
  - Something “emerges” (but the word may be rejected)
  - Something is dynamic and “self-organized”…

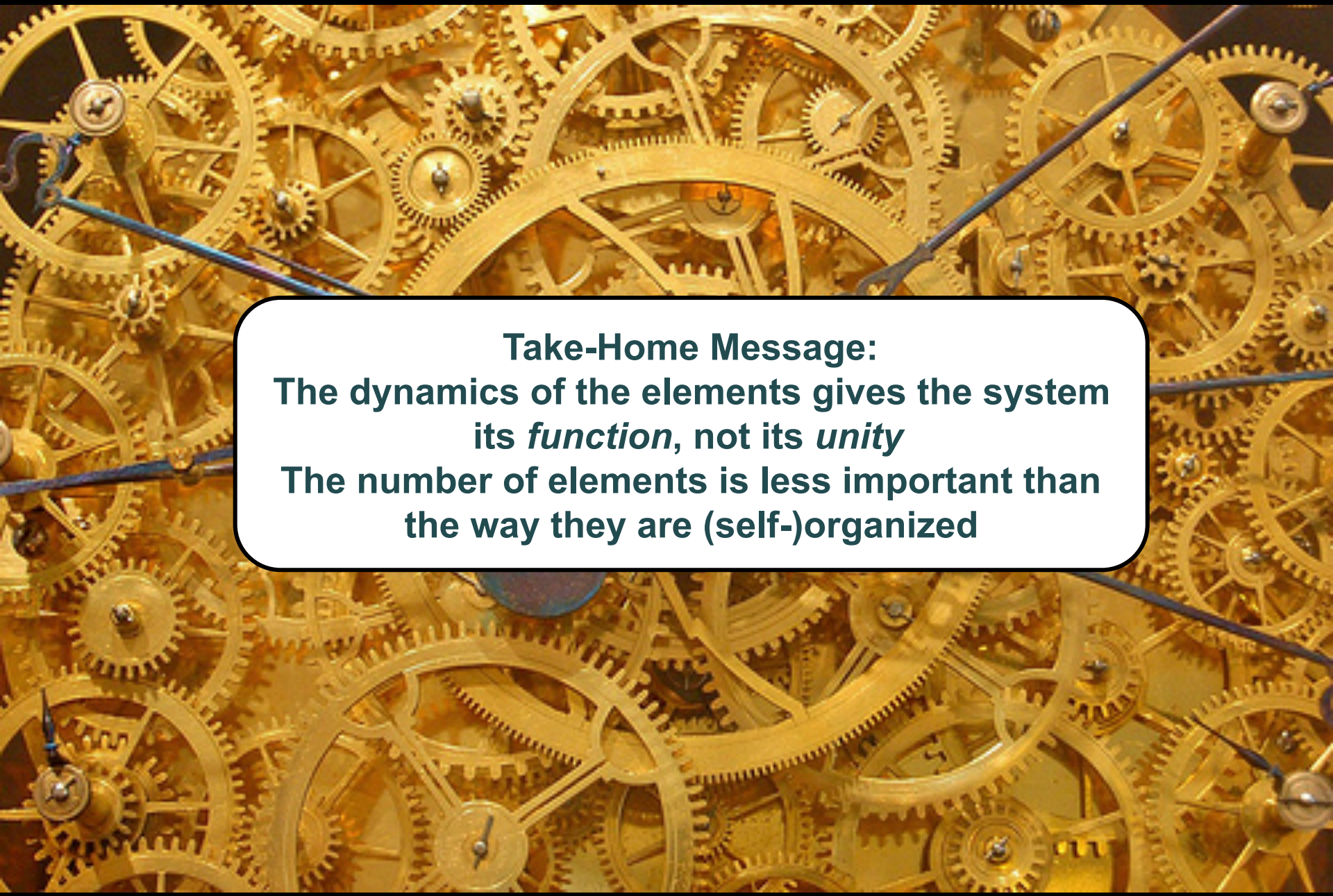
“*A system is a complex system if it is made of multiple interacting elements and if the dynamics of the interactions govern the behavior of the system, giving to it an appearance of unity from the point of view of an external observer*.”

Why this definition?

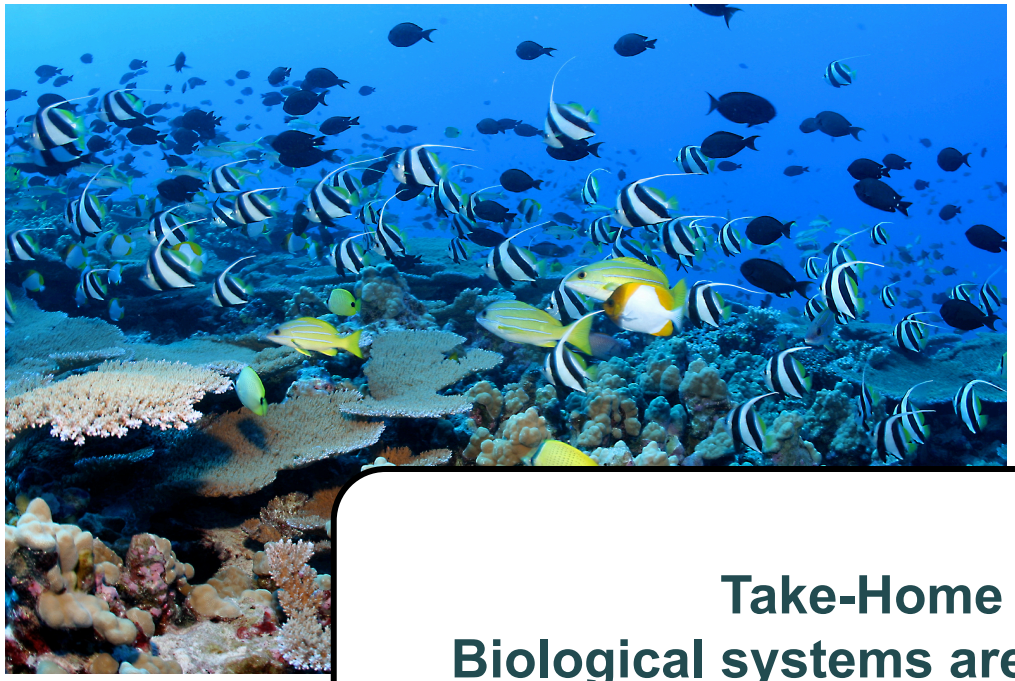
"*A system is a complex system if it is made of multiple interacting elements and if the dynamics of the interactions govern the behavior of the system, giving to it an appearance of unity from the point of view of an external observer*."

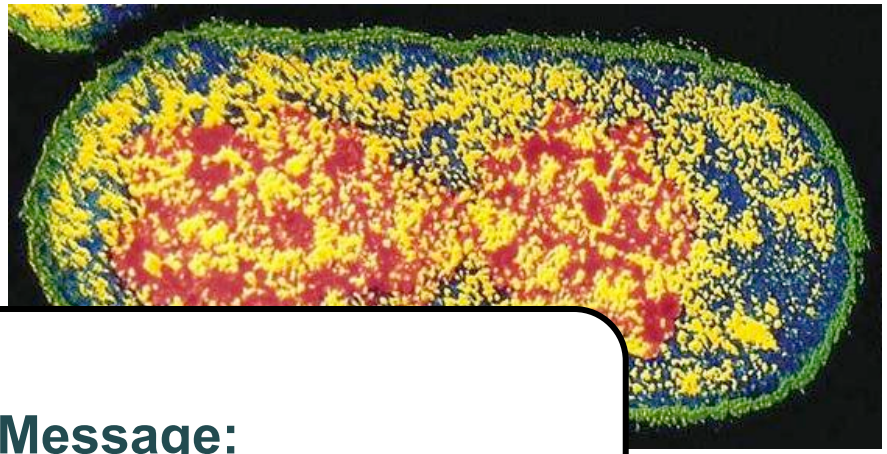
- Defines the structure of the system
  - But the structure is not enough: we also need dynamic interactions
- Subjectivity is clearly introduced
  - But it does not depend on our scientific knowledge
  - The scientist is external to the definition
- Can ask new questions (ontological/epistemological)

→ examples

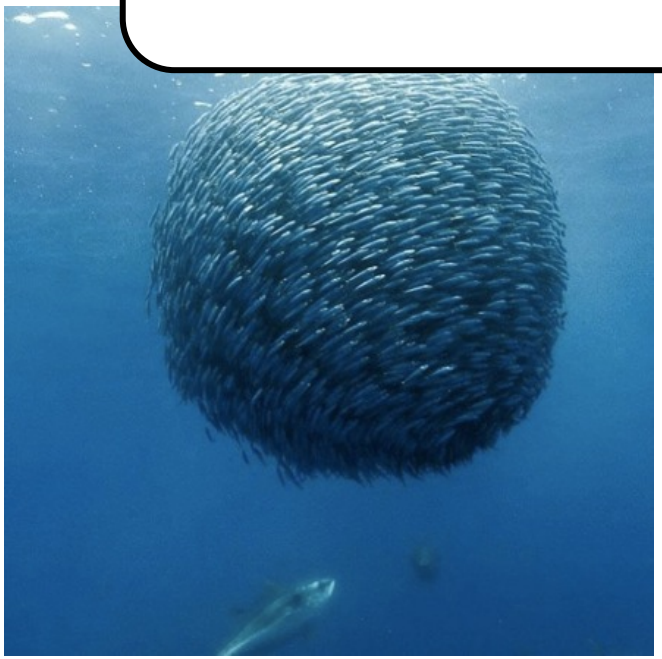


Take-Home Message:
The dynamics of the elements gives the system its *function*, not its *unity*
The number of elements is less important than the way they are (self-)organized



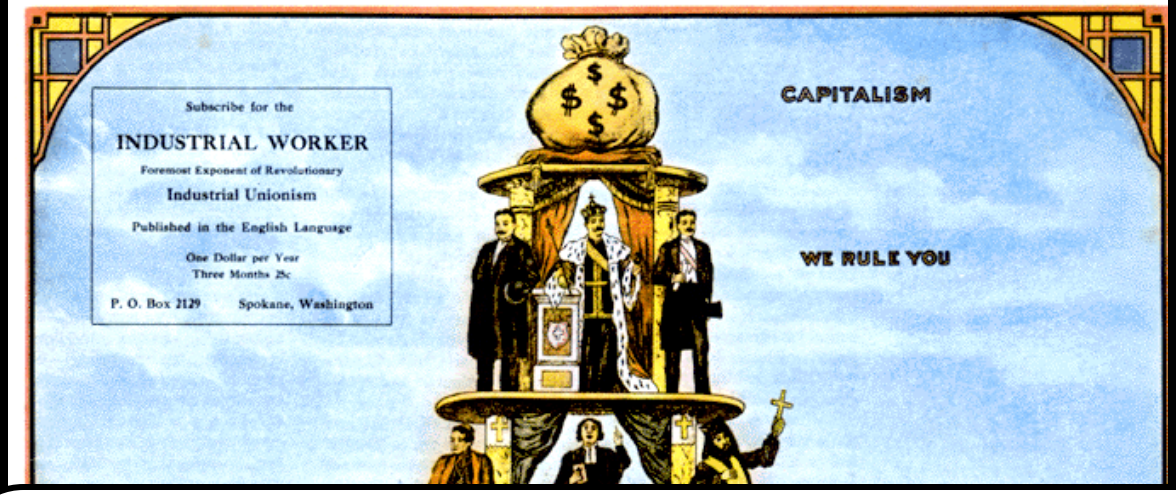


Take-Home Message:
Biological systems are always complex (?)



Take-Home Message:
The system is complex – or not – depending on the frontiers we (decide to) give it

Take-Home Message:
Social systems are always complex (?)



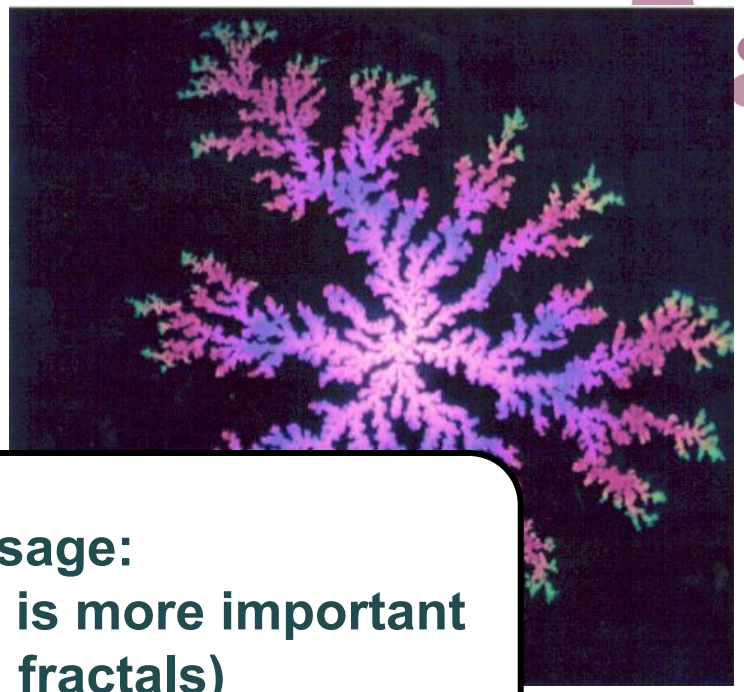
Take-Home Message:
Complex systems may (do?) not exit elsewhere than in our perception… and our perception depends on non-scientific elements (history, political opinion…)

Take-Home Message:
Complex systems can be "simple"
(Warren Weaver, 1968: disorganized complexity vs. organized complexity)

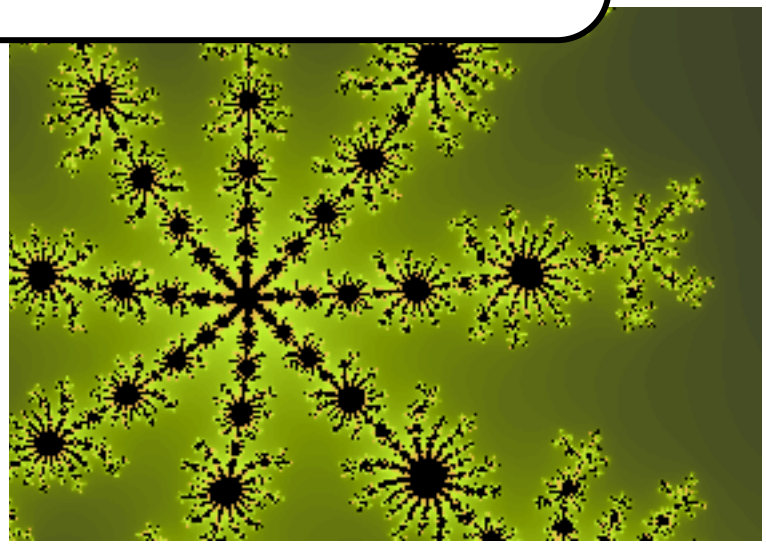
Take-Home Message:
A non-complex system can contain a complex system

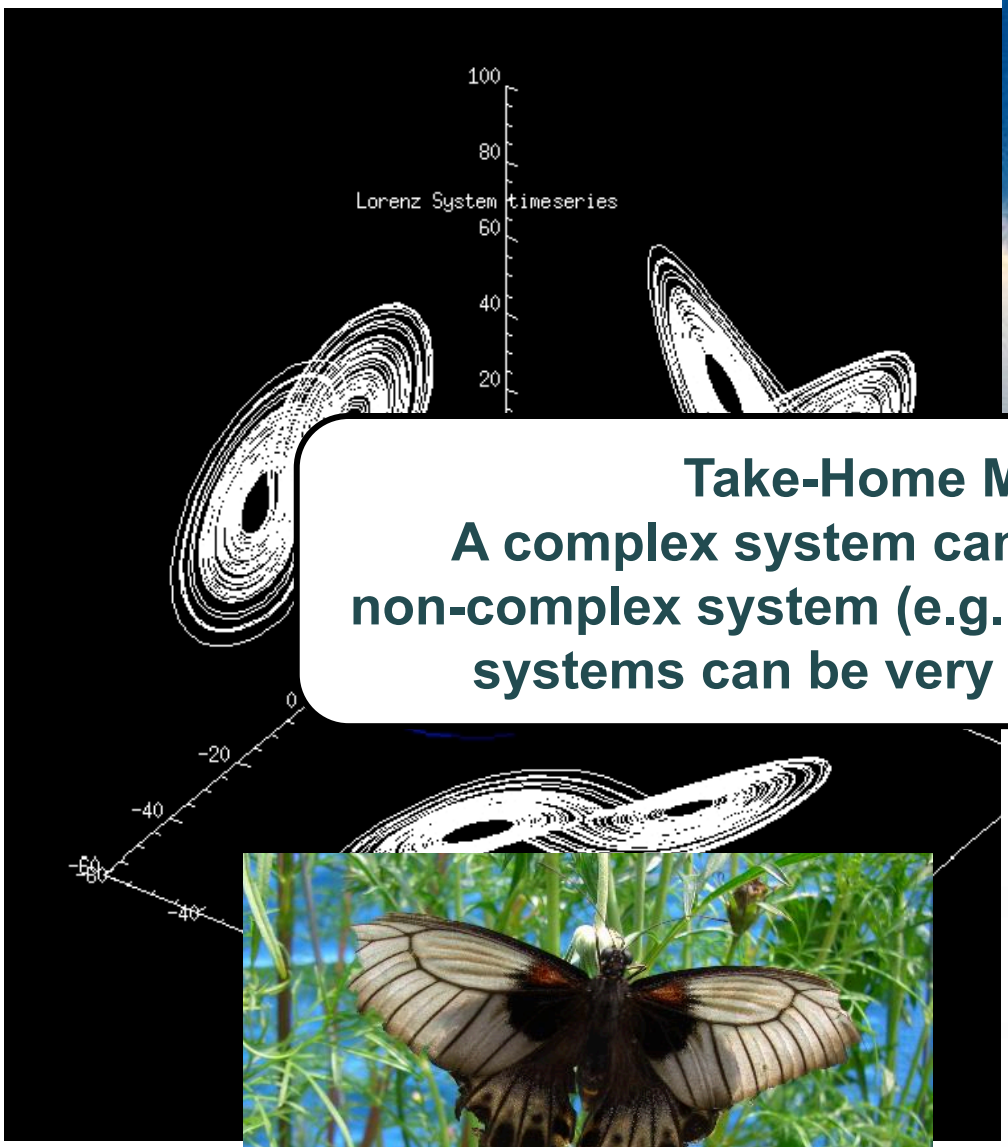
Take-Home Message:
A system can be complex – or not – depending on some of its parameters (here density)


Take-Home Message:
Engineers try to avoid complexity…
…with a mitigated success: Any (real)
system is complex at some point/scale

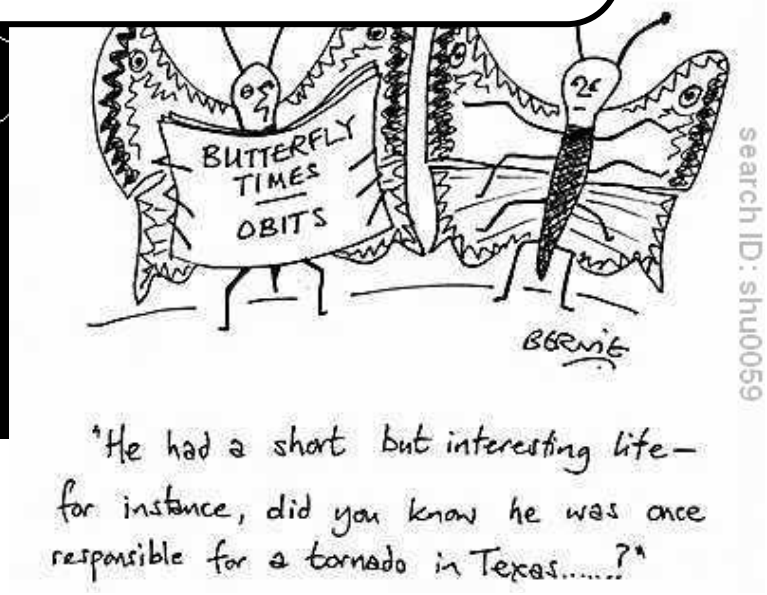


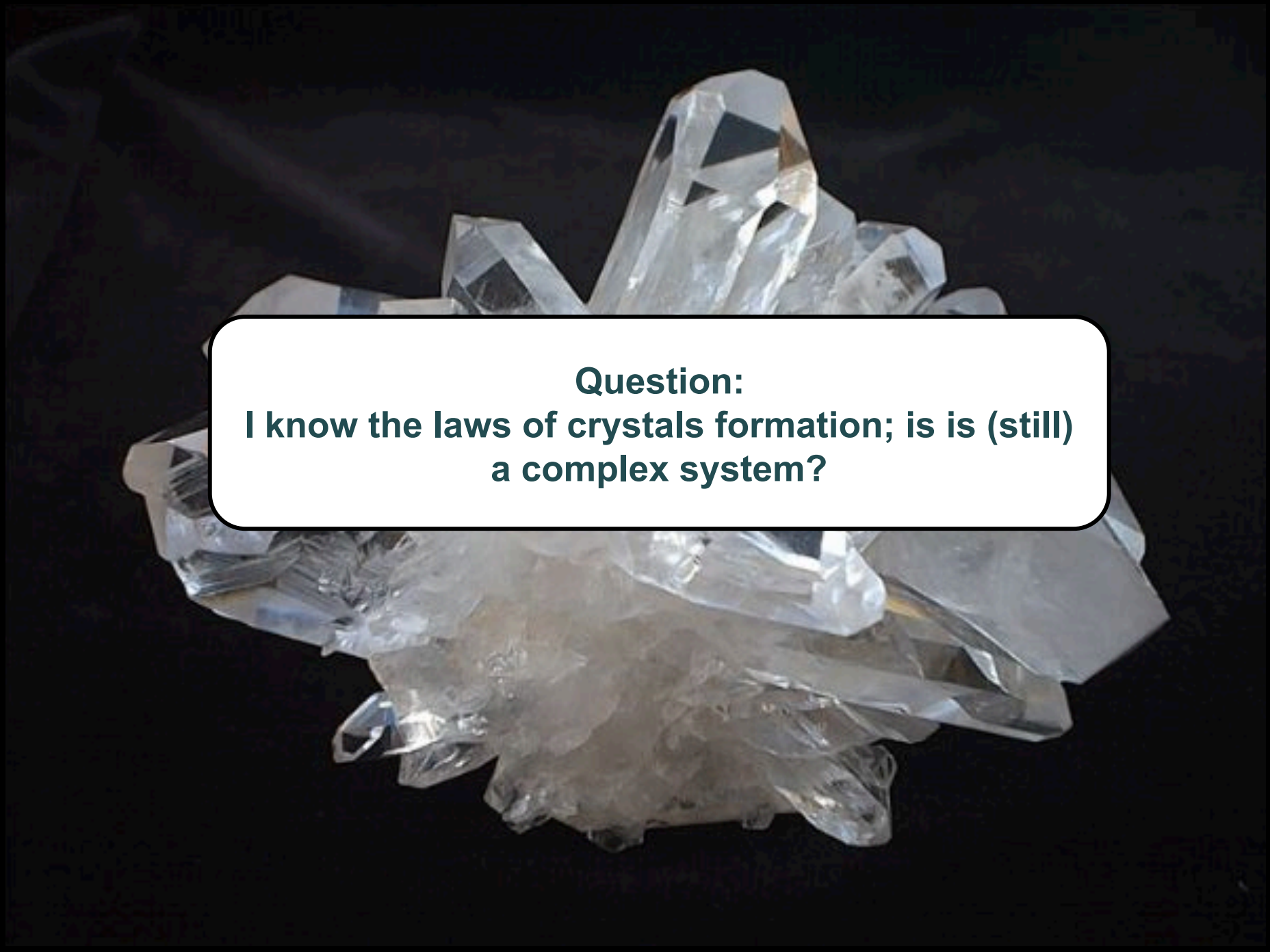
Take-Home Message:
The way a shape is produced is more important than the shape (e.g. fractals)





Take-Home Message:
A complex system can be modeled by a non-complex system (e.g. chaos); non-complex systems can be very difficult to tackle




Question:
I know the laws of crystals formation; is is (still)
a complex system?

Complex ≠ unkown

- Remember: Complex systems are ***NOT*** unexplained systems!
  - There are (fortunately) complex systems which behavior is explained
    - Any idea ?
  - There are (fortunately) open questions on systems that are NOT complex
    - Any idea ?
- Explained complex systems can be very useful to help us designing methodological principles…
  - Proof of concepts
  - Success stories

Complex ≠ complicated

- Complicated systems are “composed of a large number of elements”
- Complicated systems have “an appearance of unity from the point of view of an external observer”
- So what is the difference?
  - The unity does not come from the dynamics of the interactions
  - Complicated systems can be understood by division
- BUT: objects are not perfects …
  - Complicated objects are often *also* complex
  - Most of the work of engineers is to maintain complicated objects out of the complex regime!
- The distinction is not so clear!

Measuring complexity?

- (Quite) all the definitions of complexity are qualitative…
  - Can we give a quantitative definition of complexity?

  "*A complex system becomes more complex as the number of distinctions (distinct components, states, or aspects) and the number of relations or connections increases.*"

  [F. Heylighen, 2007]
- But these quantities are not objectively defined
  - They depends on the way you look at the system
  - The measure depends on the measurer; it is often chosen to match our subjective judgment (see next part: the C-value paradox)
- Would you say that a tiger is more alive than a bacteria?
  - Do we really need a scale of complexity?

Measuring complexity?

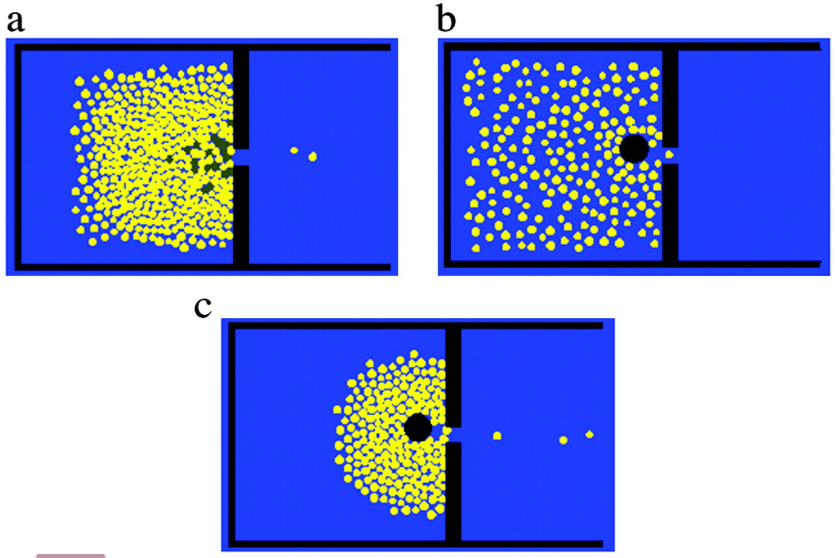
- Complexity can be measured on specific systems
  - Algorithmic complexity (time to execute a program)
  - Kolmogorov-Chaïtin complexity (size of the smallest program)
  - Bennett's logical depth (time to execute the smallest program)
- But it is not clear that what we are measuring is what we call complexity!
  - You can have classes without order… (e.g., biology)
- Actually, we are only able to measure complexity of idealized objects
  - Is complexity of the model of the object complexity of the object?
  - Most of the time
- Not sure complexity is a quantitative property!
  - Is it interesting to search for a measure? Probably yes!
  - We will learn a lot and get a deeper understanding of our field!

Why studying complex systems?

- "Because it's there" (George Mallory)
  - Science studies things that are there …

Why studying complex systems?

- "Because it's there" (George Mallory)
  - Science studies things that are there …
- Because they are prone to "natural interpretations"
  - To complex to comprehend …
  - Avoiding such natural interpretation may help to manage complex systems



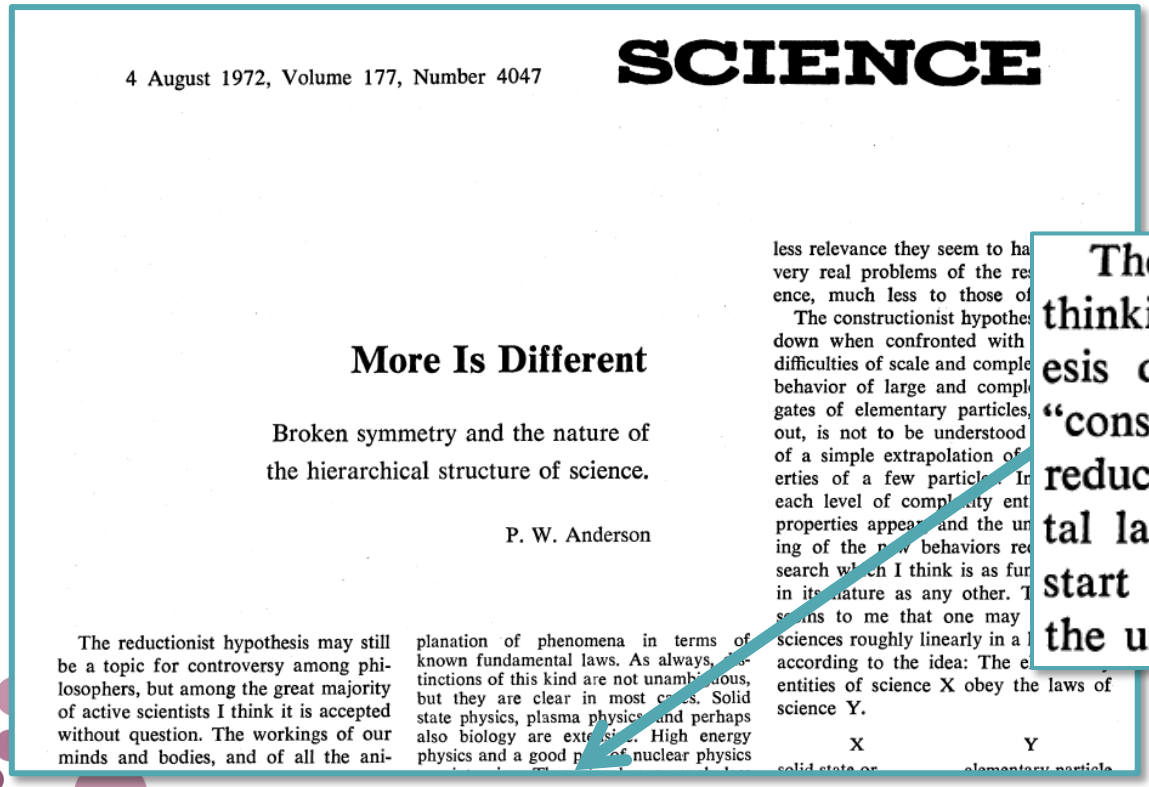
Bonabeau E PNAS 2002;99:7280-7287

Why studying complex systems?

- "Because it's there" (George Mallory)
  - Science studies things that are there …
- Because they are prone to "natural interpretations"
  - To complex to comprehend …
  - Avoiding such natural interpretation may help to manage complex systems
- Because they have properties that we would like to understand/control/create …
  - Robustness (but fragility to some specific perturbations)
  - Resilience
  - Adaptability
  - Self-organization
  - Self-X properties …

Why studying complex systems now?

- Complex systems have always existed… why is it important to study them *now*?
  1. We are reaching the limits of reductionisms



4 August 1972, Volume 177, Number 4047 **SCIENCE**

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the ani-

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are exte… High energy physics and a good p… of nuclear physics

less relevance they seem to ha… very real problems of the re… ence, much less to those of…

The constructionist hypothe… down when confronted with … difficulties of scale and comple… behavior of large and compl… gates of elementary particles, … out, is not to be understood … of a simple extrapolation of … erties of a few particle… In… each level of compl…ity ent… properties appear and the un… ing of the new behaviors re… search which I think is as fun… in its nature as any other. T… seems to me that one may … sciences roughly linearly in a … according to the idea: The e… entities of science X obey the laws of science Y.

X	Y
solid state or…	elementary particle…

> The main fallacy in this kind of thinking is that the reductionist hypothesis does not by any means imply a "constructionist" one: The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe.

Why studying complex systems now?

- Complex systems have always existed… why is it important to study them *now*?
  1. We are reaching the limits of reductionisms
  2. New data are becoming available
     - High throughput biological data (genomics, transcriptomics, ecology…)
     - Large social databases (cell phones, transport systems, mail, social networks, WoS…)
  3. Computer power is sufficient to (i) manage the new data fluxes and (2) simulate large sets of elements (model emergent behavior)
  4. Theoretical models (statistical physics, computer science) start to get out of their original field to be applied to e.g. biological systems or social science

The science of complex systems

- We can (more or less) define a "complex system" but what is the "science of complex systems"
  - ~any system is a complex system at some levels of description
  - Does it implies that "science of complex systems = science"?
  - Hope you'll agree that it is absurd! (at least)
- How can you define a science?
  - E.g., Biology, Chemistry and Physics are all working on DNA
  - A science is not defined by its objects but rather by its questions
- The science of complex systems is ***NOT*** the science of complex objects ***NOR*** the science of complex questions!
  - It is the science of questions that are specific to complex systems

Back to the definition

"A system is a complex system if it is made of a large number of interacting elements and if the dynamics of these interactions govern the behavior of the system, giving to it an appearance of unity from the point of view of an external observer."

Back to the definition

"A system is a complex system if it is made of a ***large number of interacting elements*** and if the dynamics of these interactions govern the behavior of the system, giving to it ***an appearance of unity*** from the point of view of an external observer."

- So the question is:
  - Given the elements and their interactions, how can we quantify/ understand/reproduce the appearance of unity?
- From this general question, we can derive
  - More specific questions (field specific, intermediate)
  - Differences from other fields (possibly looking at the same object, e.g., molecular biology vs. systems biology)
  - The embryo of a methodology to study complex systems…

The science of complex systems

- Different kind of research can be done in the context of complex systems
  - Data-driven research (real systems, large data-bases)
  - Theoretical research (identify general features)
  - Modeling and simulation (capture emergence)
- Objectives:
  - Predict
  - Control
  - Design
- A science intrinsically interdisciplinary…
  - The real difficulty in CSS is interdisciplinarity…

Interdisciplinarity

- Interdisciplinarity is often promoted
  - But rarely realized (powerful and dangerous)
- Crossing disciplines boundaries is a difficult exercise that needs **time** and **tact**…
  - Be modest: All scientific disciplines have a long history
  - Be open-minded: All scientific disciplines have their own habits
  - Be honest: What do you want to show? (where do you want to publish? To whom to you want to explain your results?)
  - **Never suppose you can provoke a scientific revolution from the outside!**

"The burden of proof [in alife] is on us to explain our results to biologists in their own language and in their our journals"

[Miller, 1995]

Methodology

- Applied CSS: Three main questions can be derived from the central one:
  1. Description: What is the "unity" of the system? What are the elements? How can we describe both levels accurately?
  2. Understanding: What is the link between the dynamic of local interactions and the unity of the global system?
  3. Why do we perceive this system as a "unity"? What is "emergence"? (coping with the subjective part of the definition)
- Theoretical CSS: Can we identify universal features (laws) in complex systems
  1. Toy models: understand the laws
  2. Formal models: prove the laws
  3. Measures and tools to cope with a large amount of data

Research in complex systems (1): describe the system

- Build a "complete" description of the system
  - At the local level (what are the elements?)
  - At the global level (what is the unity?)
- Remember that you can miss important points! To draw a "complete" description, you MUST be very careful…
  - You must have a very good knowledge of the system…

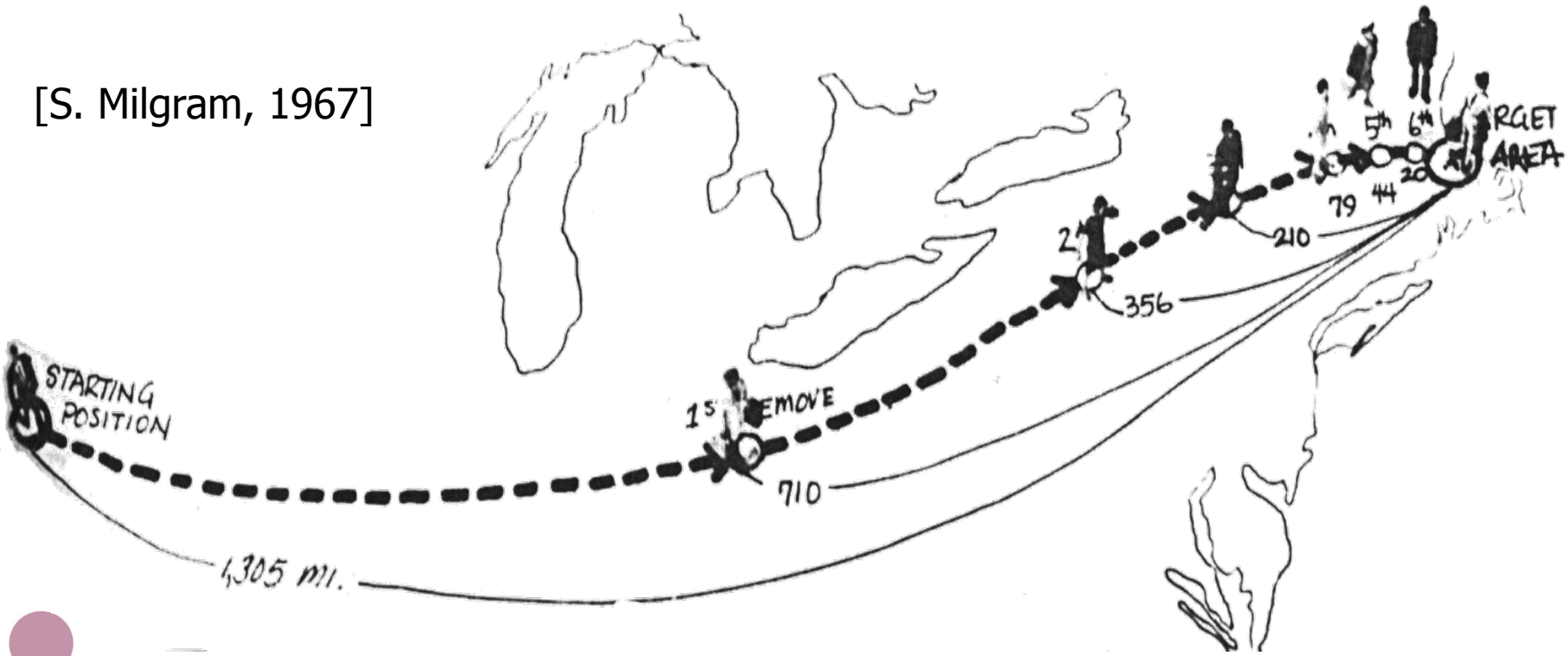
    → Back to the problem of interdisciplinarity…
- Remember that a description is always
  - A subjective selection (of properties, scales, frontiers,…)
  - Dependent on the point of view (position, scale, time, science…)

    → Back to the problem of interdisciplinarity!

Research in complex systems (1): describe the system

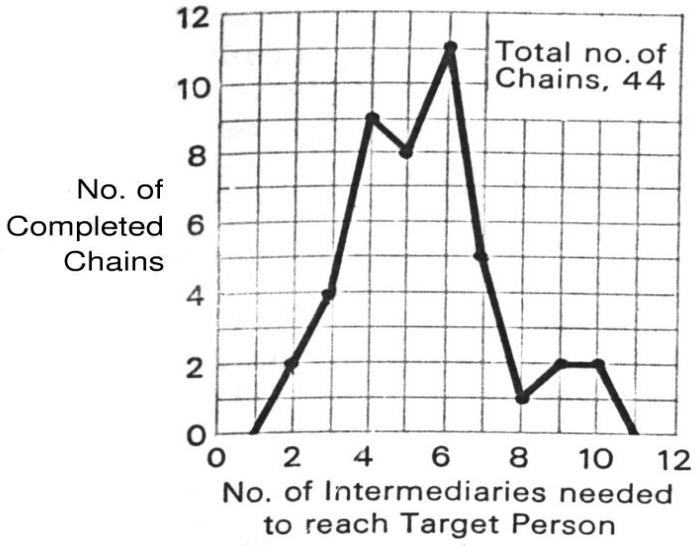
- The example of complex (social) networks

[S. Milgram, 1967]



Research in complex systems (1): describe the system

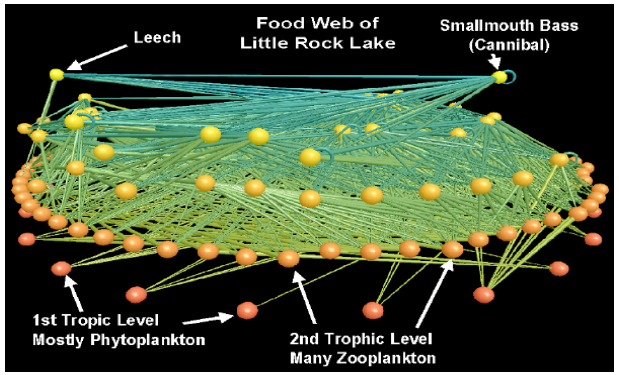
- The example of complex (social) networks
- Average distance in North-America:
  - $l \approx 6$
  - "six degree of separation"
- Classical distances in networks
  - $l \sim \log N$ ("small world")
  - $l \sim N$ ("grid")

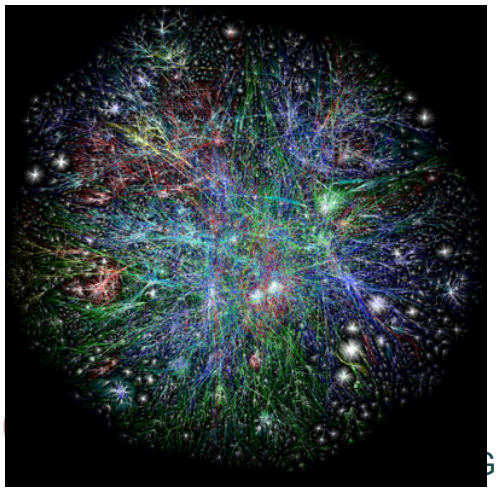


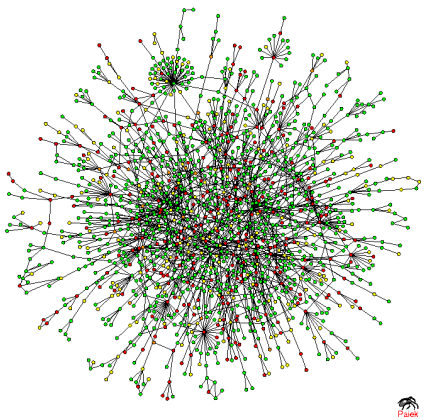
In the Nebraska Study the chains varied from two to 10 intermediate acquaintances with the median at five.

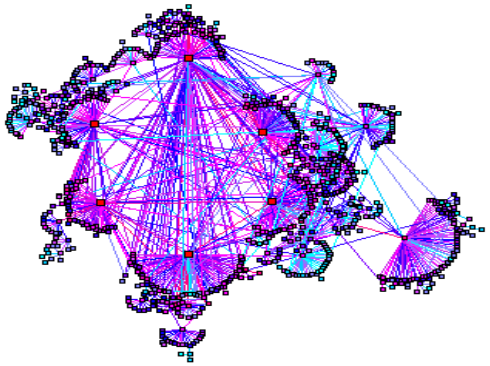
Research in complex systems (1): describe the system

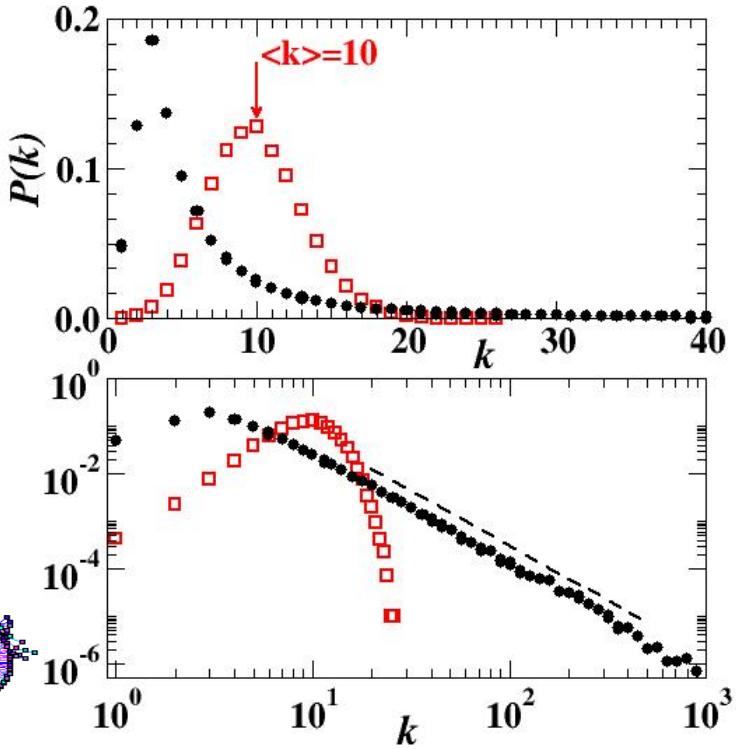
- Lots of real networks are scale-free and small world …









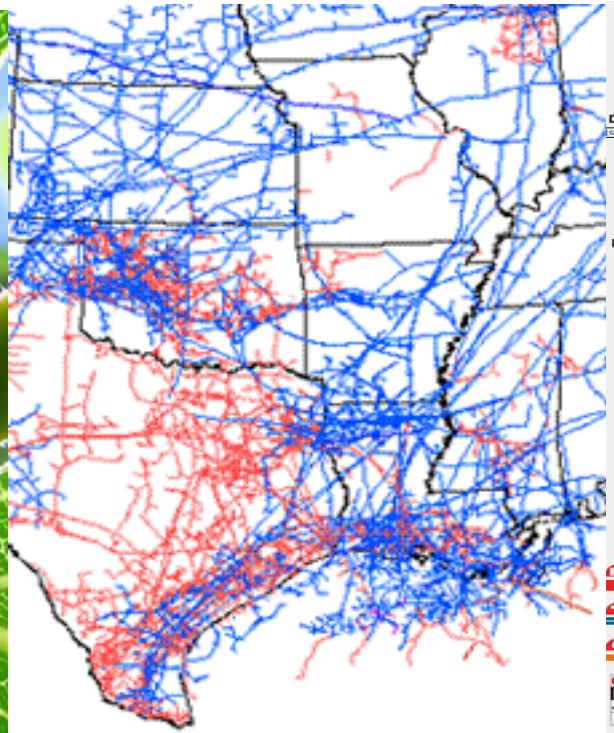


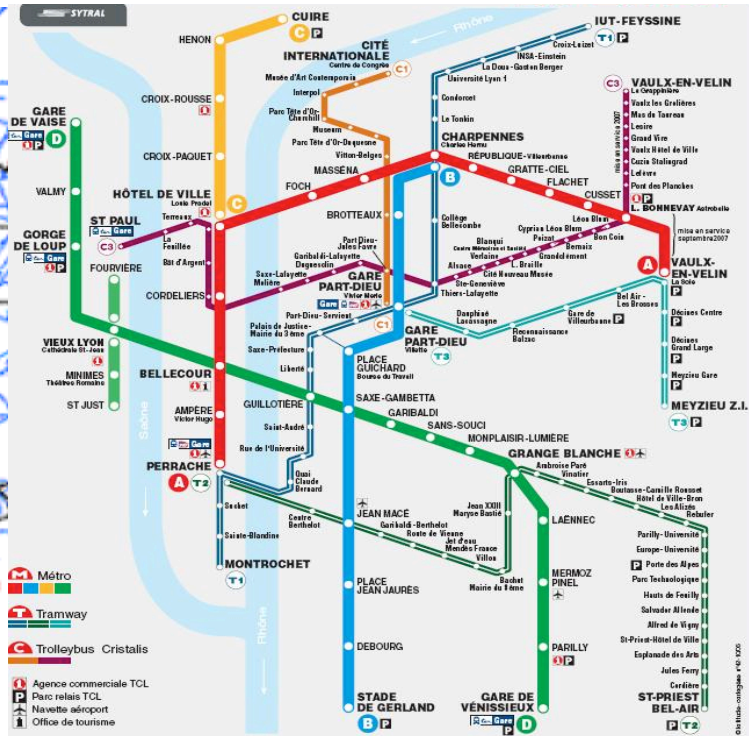
Research in complex systems (1): describe the system

- Once you have collected the data, you can browse the description and seek for hidden structures
  - E.g., in the context of social networks
    - Community detection
    - Invariant characteristics
    - Scale laws, hubs,…
- Then you can search for an origin of these structures
  - Switch to question 2 …
  - Note that the origin of the structures may not be “complex” !

Research in complex systems (2): Understanding

- Example of complex (social) networks (continued)
  - Networks share (or not) some common principles (e.g. branching)
  - Can these principles explain the global structure?





Research in complex systems (2): Understanding

- Probably the main goal!
- How can we do?
  - Complex systems are often difficult to tackle (either experimentally or by though experiments)
  - Dynamic non-linear interactions may contradict our intuition
  - We need tools to explore the behavior of the system and to explain the emergent comportment
- We need “models” to help us to understand the behavior of the system…
  - How can models help us to understand something?
  - How to use models in science?

What is a model?

"To an observer *B*, an object *A is a model of an object *A* to the extent that *B* can use *A** to answer questions that interest him about *A*."**

(Marvin Minsky, 1995)

- The "model triad": To be a model, an object (system, equations, software, …) needs to be linked to a triplet of elements:
  - <Objet, Question, Observer>
  - The observer is the link between model domain to object domain
- The model must be used to answer questions: it is a scientific instrument
  - It must be used like an instrument (experimental method)
  - But models depend on the question more than on the object
  - It is not an instrument "like the others" (e.g. a microscope)

How to use models?

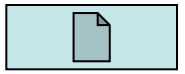
- The difficulty of modeling is actually a difficulty of inter-disciplinarity
  - **You** (as a modeler) must be ready to discuss with *biologists* to co[illegible]rrect
- Implic[illegible]

> **"The decisive thing with modeling is not the model per se, but what the model and working with the model does to our mind. […] It could be argued that a criterion to determine good models is that they are no longer needed afterward"**
>
> **(V. Grimm, 1999)**

1. pi[illegible]
2. co[illegible]
3. do[illegible]
4. develop a simulation that extends current models (not necessarily a beautiful "looking like" simulation)
5. use simulation experiments to explore cause and effect and follow a scientific method (repeats, stat., lab. book,…)
6. publish in real peer-reviewed journals of the field

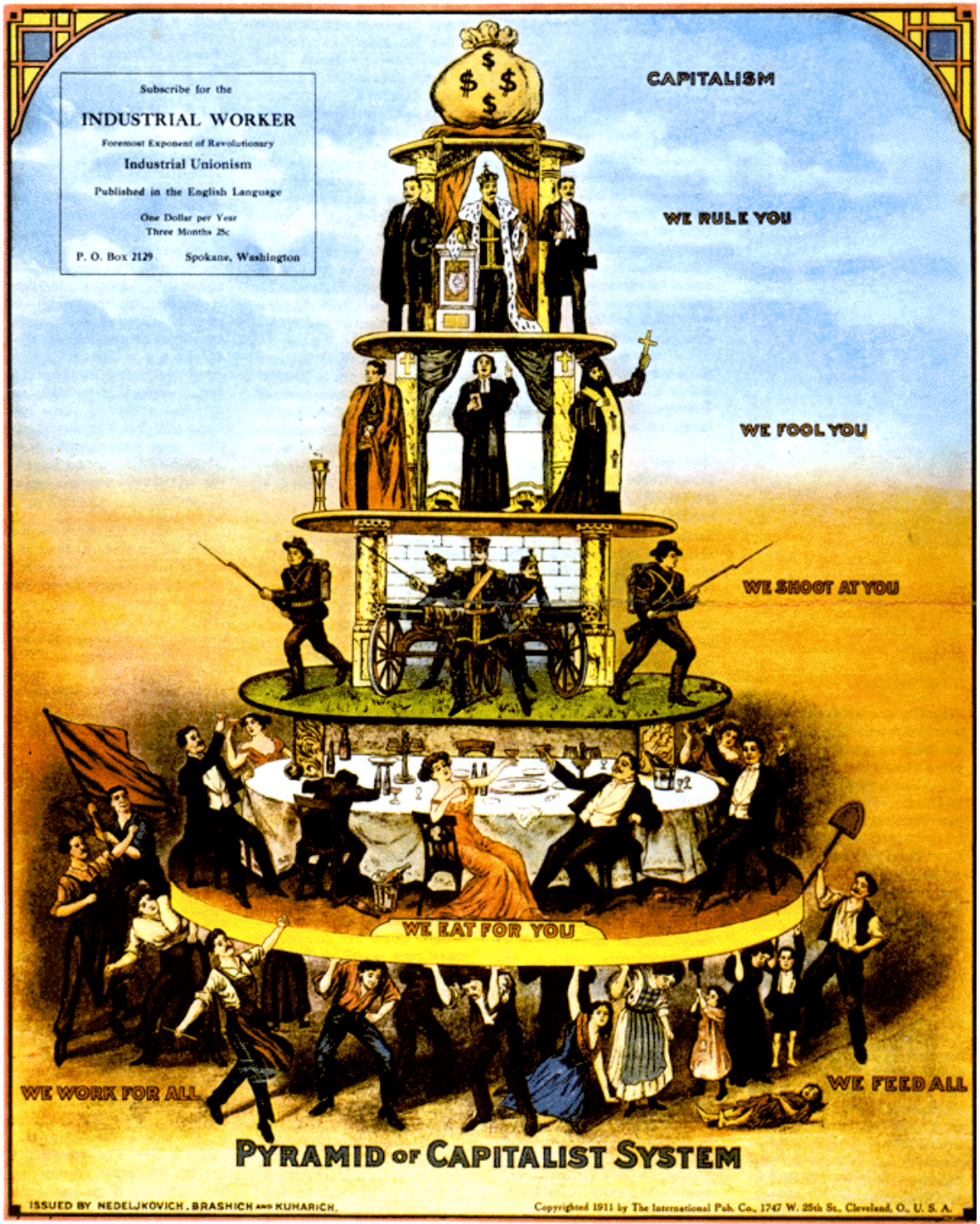
Research in complex systems (3) : cognitive science?

- "[…] ***an appearance of unity*** from the point of view of an external observer"
  - Why do we see a unity there but not there?
  - What are the limits of the system?
  - A trivial example : the game of life



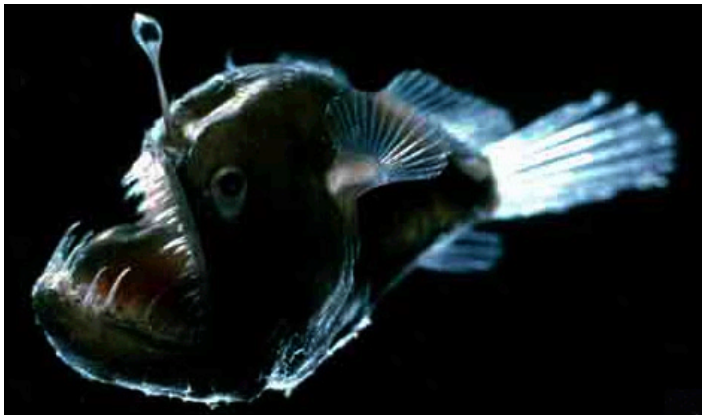
- Can the science of complex systems help us to understand how our brain work?
  - Our brain actually is a complex system (but this is question 2!)
  - Our brain is the product of evolution; it's structure and behavior have been selected because they efficiently categorize the world… that is a complex system!
- Difficult epistemological questions (what is emergence? What is a frontier?)


steinmann


Subscribe for the
INDUSTRIAL WORKER
Foremost Exponent of Revolutionary
Industrial Unionism
Published in the English Language
One Dollar per Year
Three Months 25c
P. O. Box 2129 Spokane, Washington
CAPITALISM
WE RULE YOU
WE FOOL YOU
WE SHOOT AT YOU
WE EAT FOR YOU
WE WORK FOR ALL
WE FEED ALL
PYRAMID OF CAPITALIST SYSTEM
ISSUED BY NEDELJKOVICH, BRASHICH AND KUHARICH.
Copyrighted 1911 by The International Pub. Co., 1747 W. 25th St., Cleveland, O., U. S. A.

The notion of (complex) system questions our classifications

- Thinking “complex” often means crossing the frontiers of the considered system…
  - E.g., organism/ecosystems
  - E.g., petri dish/human gut
  - E.g., web site/web network



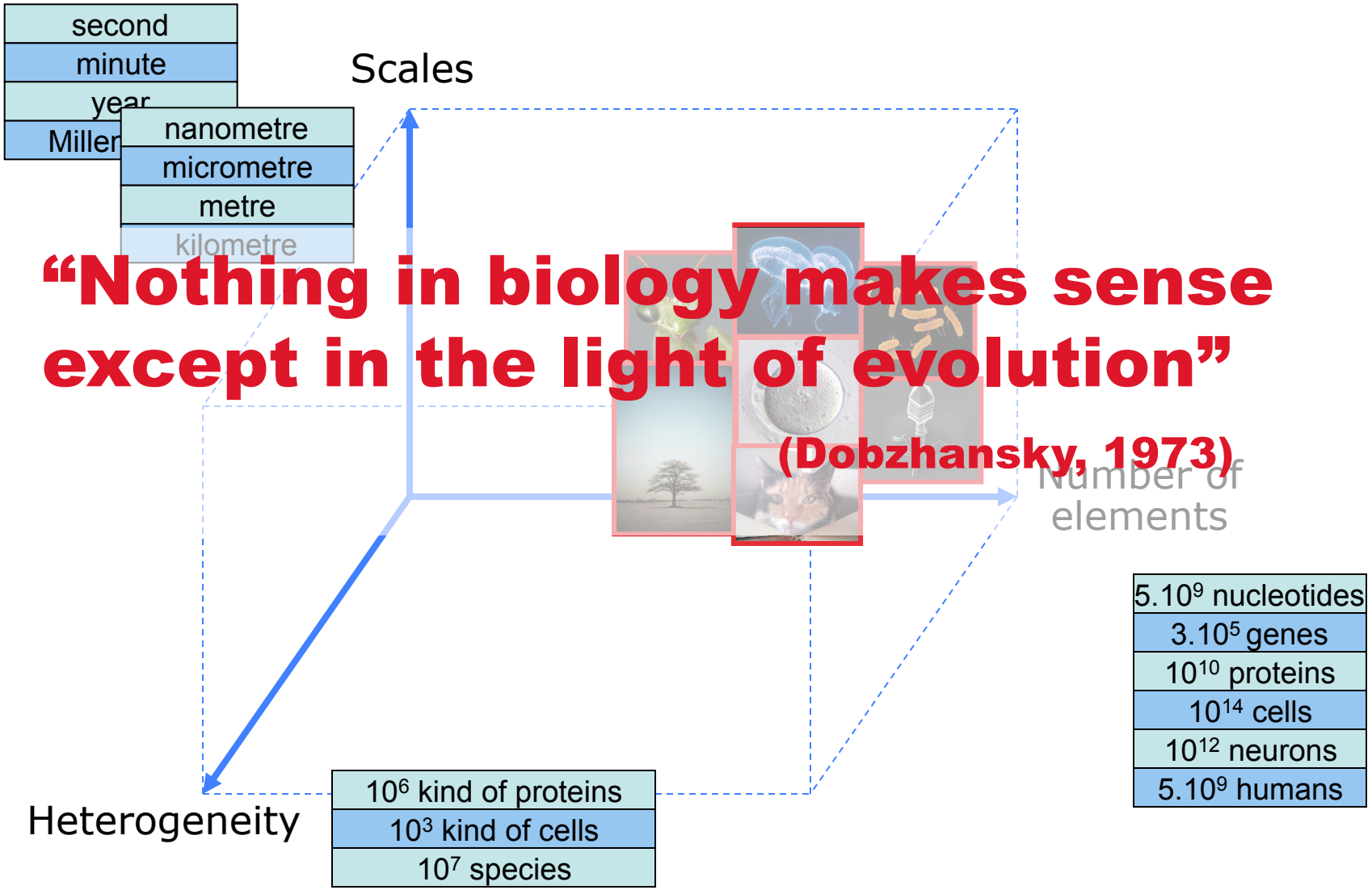
Complex Systems Summer School

Part 2: Digital genetics, a view on the origin of biological complexity

Guillaume Beslon

INSA – LIRIS – IXXI

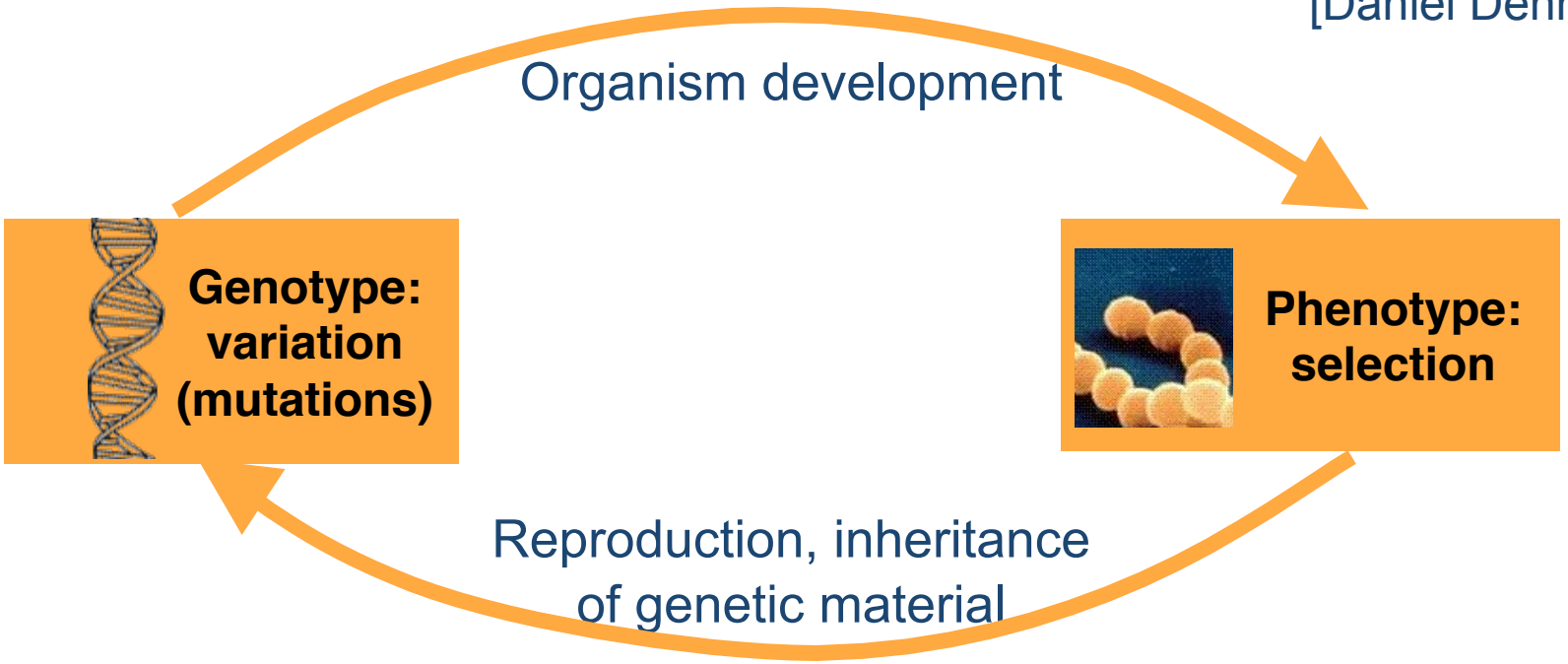
Biocomplexity?



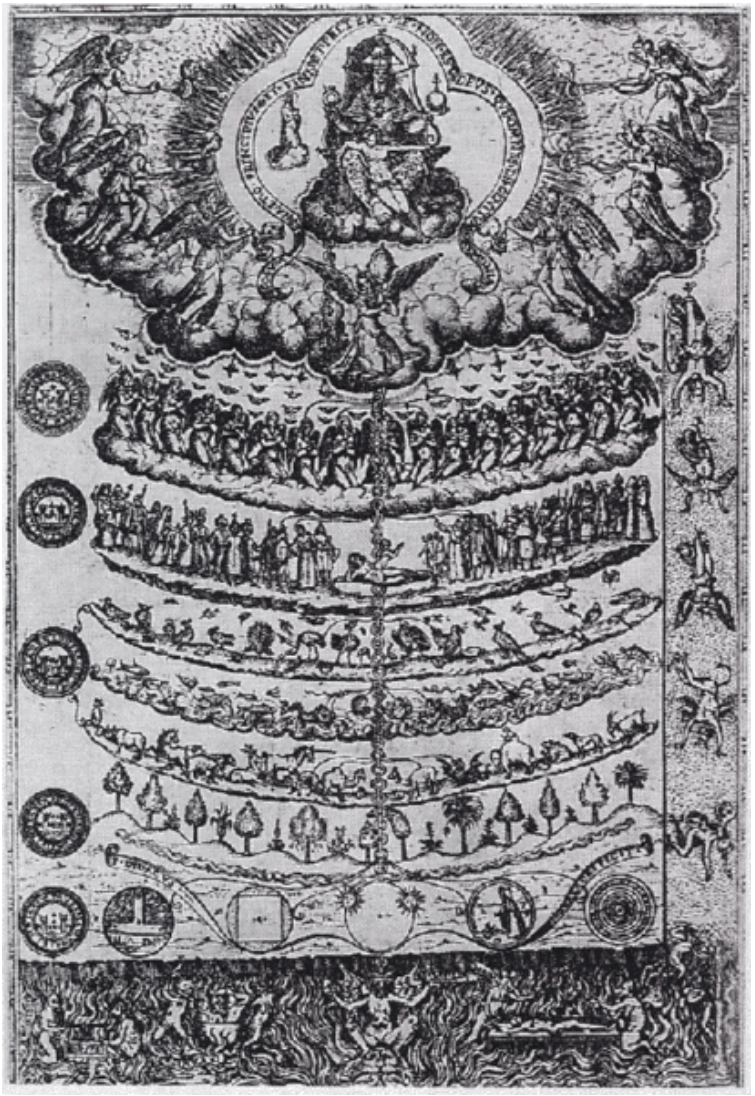
Evolution?

"*Evolution will occur whenever and wherever three conditions are met: replication, variation (mutation), and differential fitness (competition).*"

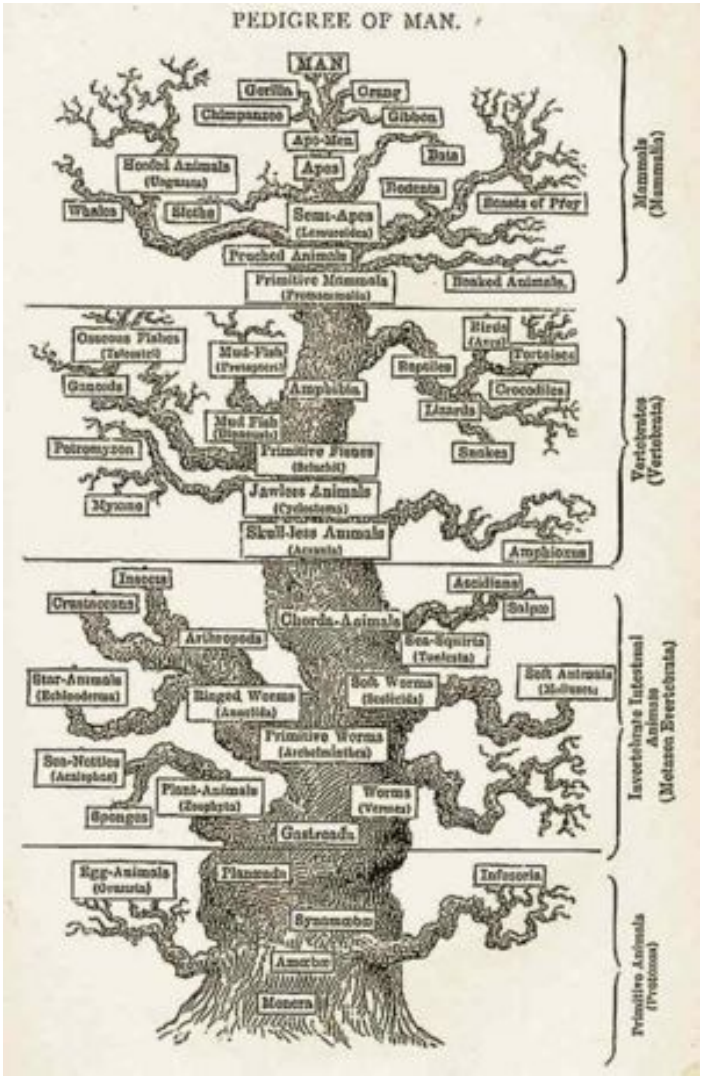
[Daniel Dennett]



Evolution “creates” complex systems



Before Darwin



After Darwin

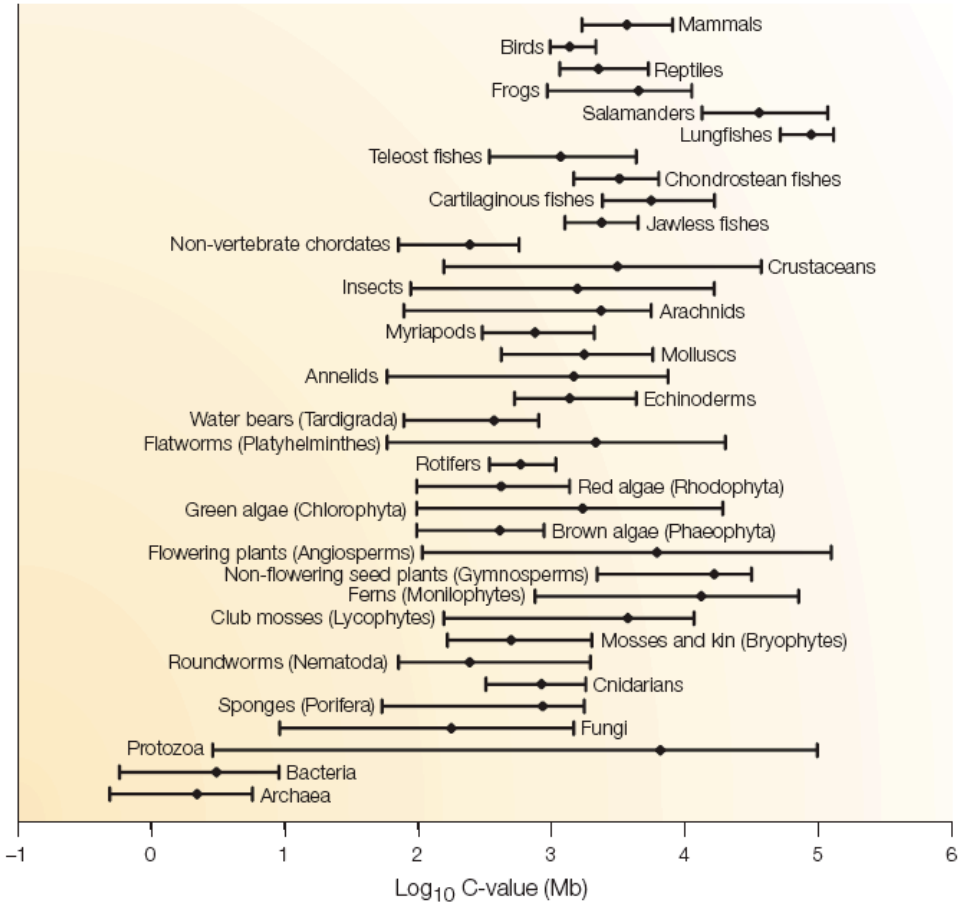
But what are the rules?

C-value paradox: genome size does not correlate with organismal *Complexity*

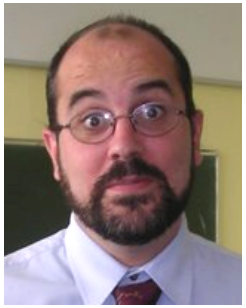
Gregory, G. T. (2005) Synergy between sequence and size ir large-scale genomics. *Nat. Rev. Genet.*,6(9): 699-708

Box 1 | **Extensive variation in genome size within and among the main groups of life**

Ever since the first general surveys of nuclear DNA content were carried out in the early 1950s it has been apparent that eukaryotic genome sizes vary enormously and that this is unrelated to intuitive ideas of morphological complexity[2]. This discrepancy between genome size and complexity remains clear more than half a century later, with genome sizes now available for nearly 9,000 species of animals and plants[10,11]. In prokaryotes, genome size and gene number are strongly correlated[86], but in eukaryotes the vast majority of nuclear DNA is non-coding (FIG. 1; BOX 3). Nevertheless, there is some overlap in genome size between the largest bacteria and the smallest parasitic protists. The figure illustrates the means and overall ranges of genome size that have been observed so far in the main groups of living organisms, and are loosely arranged according to common ideas of complexity to further emphasize the disparity between this parameter and genome size. Some commonly cited extreme values for amoebae (700,000 Mb) have been omitted, as there is considerable uncertainty about the accuracy of these measurements and the ploidy level of the species involved[10,87].

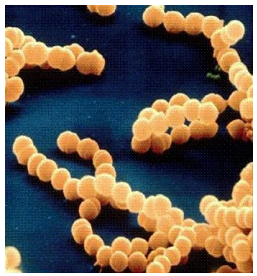


First step: description



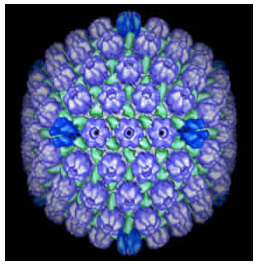
Homo sapiens

~3 billions bp
~25 000 genes



Neisseria meningitidis

~2 millions bp
~2 000 genes



Herpes HSV-1

~150 000 bp
~100 genes

Description

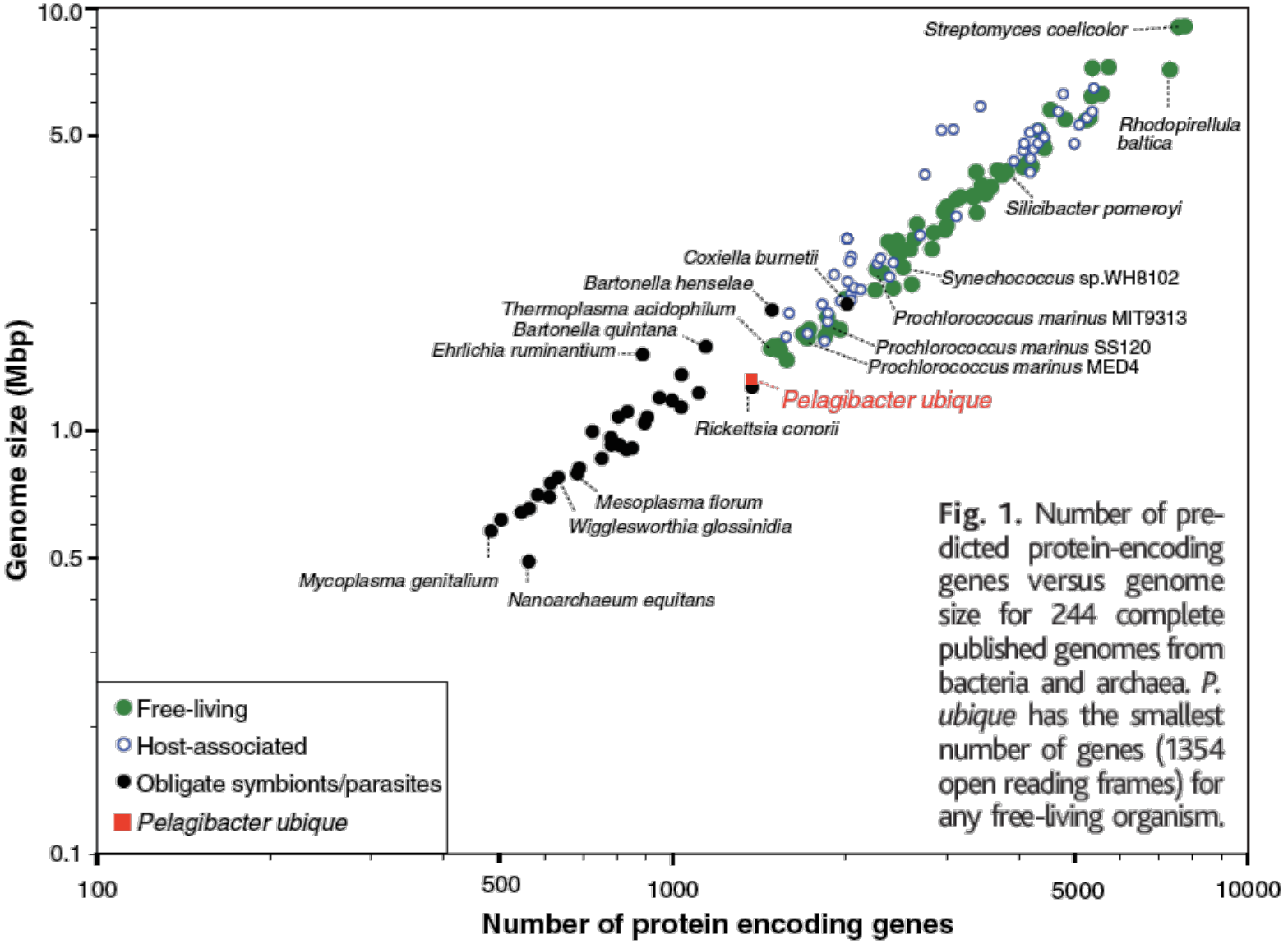


Fig. 1. Number of predicted protein-encoding genes versus genome size for 244 complete published genomes from bacteria and archaea. *P. ubique* has the smallest number of genes (1354 open reading frames) for any free-living organism.

Giovannoni, S. J., *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* (2005) vol. 309 (5738) pp. 1242-5

Second step: modeling

- Numerous regularities/irregularities among organisms (“unity”)
  - C-Value paradox and genome structures
  - Networks structure
  - Allometric scaling laws
- These regularities all come from evolution but the evolutionary causes can be very different
  - Selection (the structure provides a selective advantage)
  - Founding effect (the common ancestor was already like that)
  - Mutational effect (the nature of the mutational process creates that)
  - Drift (accumulation of neutral/quasi-neutral mutations)
  - Indirect selection (the organisms are likely to evolve better)

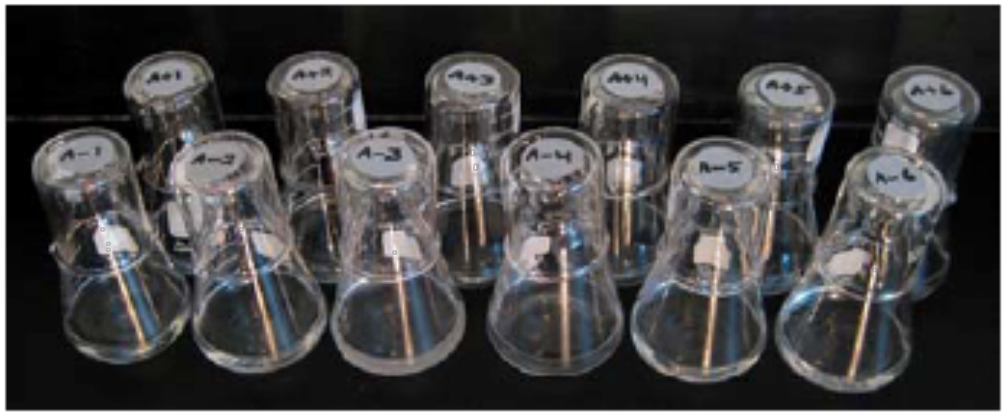
→How to decipher the roots of biological complexity?

(remember that we have to start with a question!)

in vivo experimental evolution

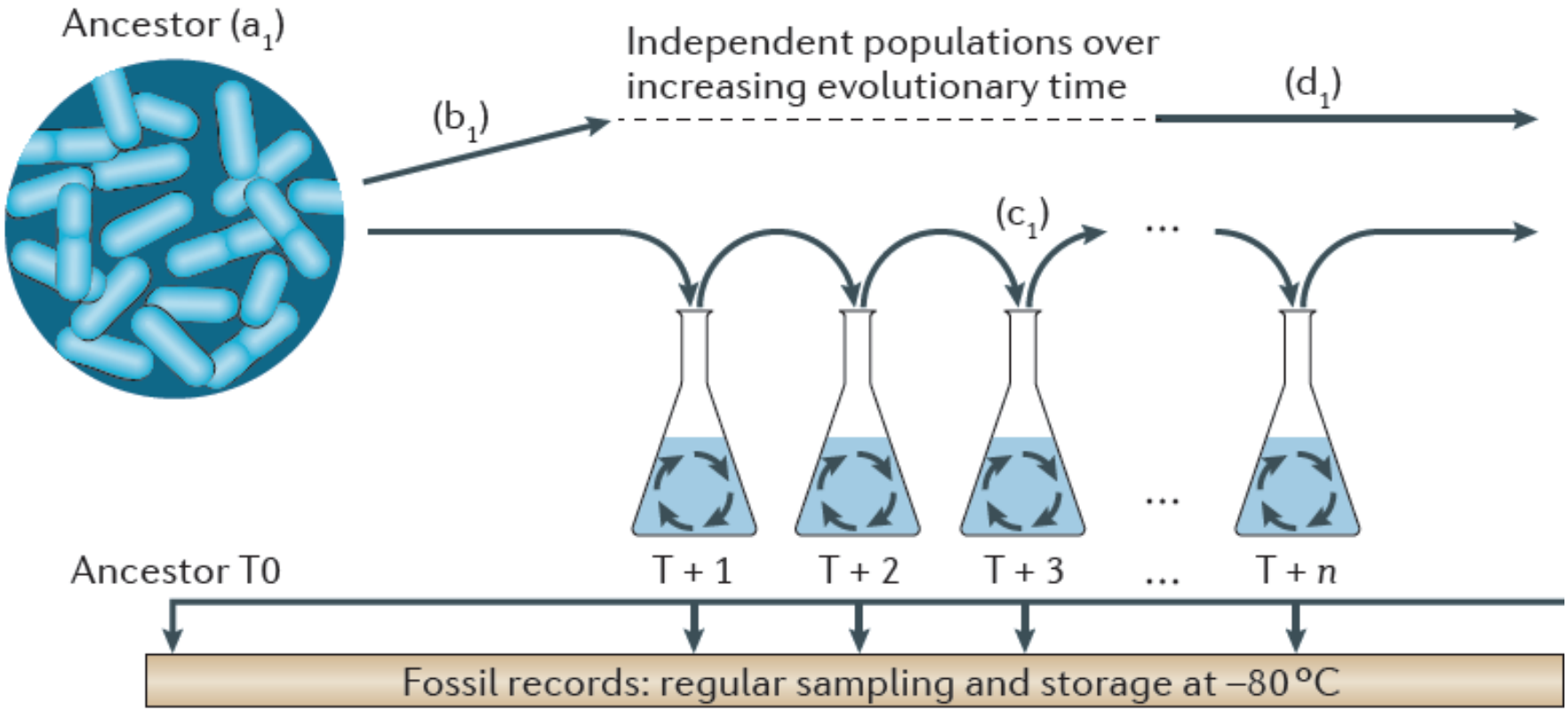
- Experimental approaches are often considered impossible in evolution
- Yet, evolution experiments can be performed with organisms
  - Cheep, small and abundant,
  - Easy to culture in controllable environments
  - Short replication time
  - Measurable and well known (model organisms)
  - Easy to freeze and revive

→ Virus and phages, bacteria (*E. coli*, *salmonella*, …), unicellular eukaryotes (yeast) or multicellular "simples" ones (*C. elegans*, drosophila)



R. Lenski Lab.
(Michigan State Univ.)
LTEE with *E. coli*
1988-2012
~45 000 generations

in vivo experimental evolution



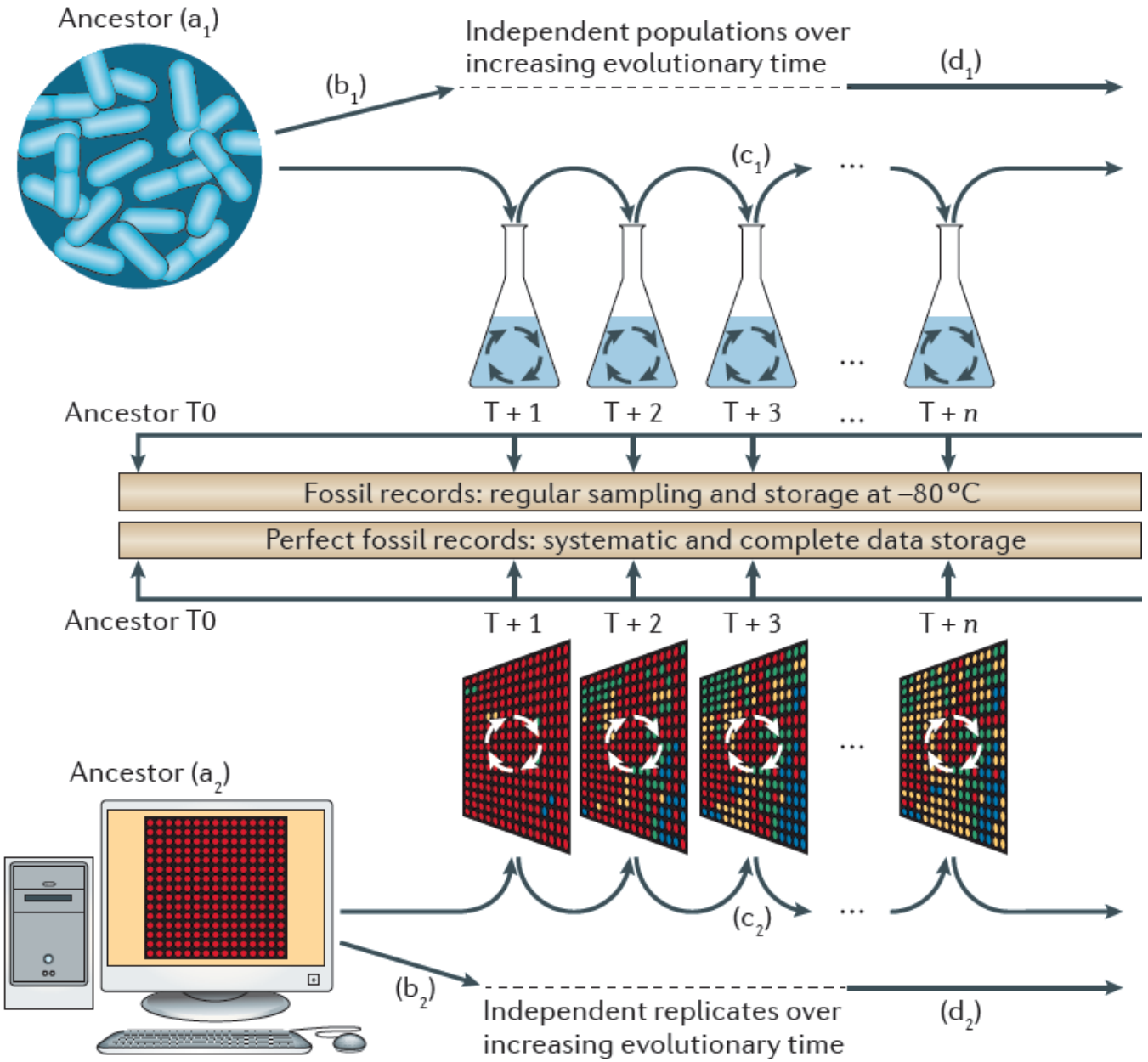
Experimental evolution is not enough

- All known organisms share parts of their evolutionary history
  - We all come from LUCA (~3.5 billion years ago)
- Conditions are always changed by the experimental setup
  - What are the consequences on the evolutionary process?
- How can we analyze the results?
  - Real organisms are too complex for us!

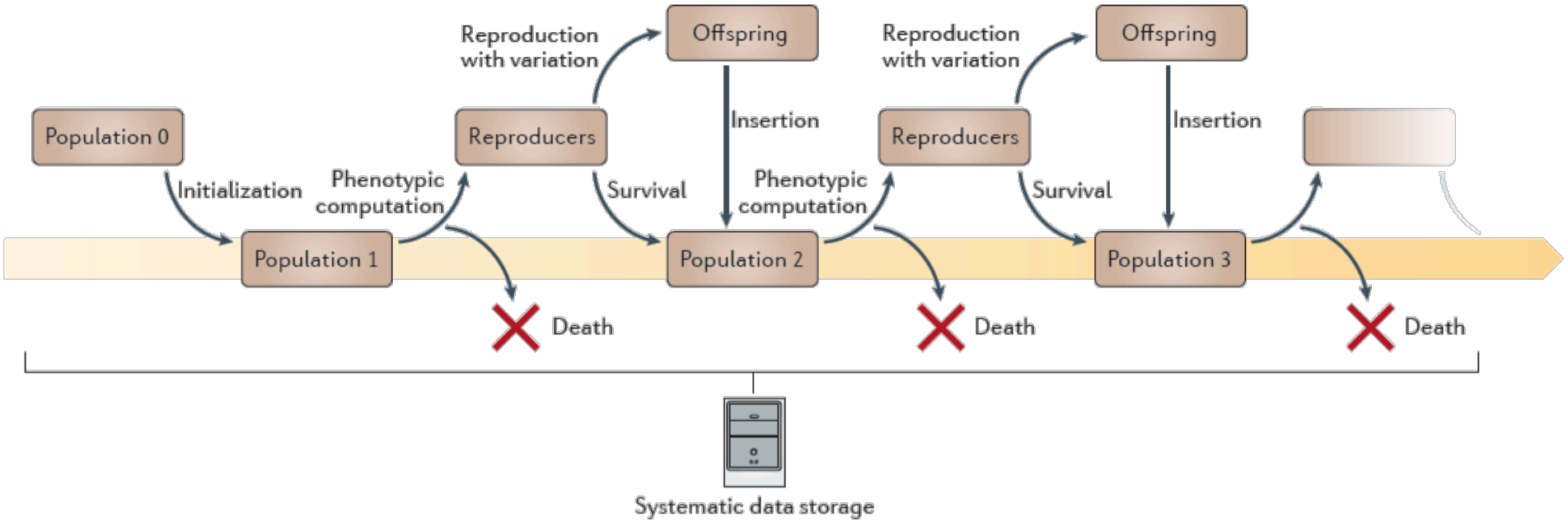
"*So far, we have been able to study only one evolving system and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we have to look at artificial ones.*"

[John Maynard Smith, 1992]

→ "digital genetics"


Ancestor (a_1)
Independent populations over increasing evolutionary time
(b_1)
(d_1)
(c_1)
Ancestor T0
T + 1
T + 2
T + 3
T + n
Fossil records: regular sampling and storage at –80 °C
Perfect fossil records: systematic and complete data storage
Ancestor T0
T + 1
T + 2
T + 3
T + n
Ancestor (a_2)
(c_2)
(b_2)
Independent replicates over increasing evolutionary time
(d_2)

The evolutionary process is modeled in three steps



“digital organisms”

Table 2 | **Genome formalisms in *in silico* experimental evolution**

Formalism*	Description	Questions addressed
Program	The genome is a sequence of instructions in a ...aming language. The fitness of the program depends ... ability to create copies of itself in the computer's memo... .../or to perform specific computations	...e emergence of parasites and hyperparasites[38] ...e evolution of robustness, evolvability, complexity and ...odularity[3,87,142,147,150] ...e adaptive radiation of species[84] ...e information threshold (the maximum amount of ...formation that can be evolutionarily maintained)[151]
Allelic	The genome is made up of a fixed gene numbe... ...ach gene can exist in a finite or infinite number of alleles; all... ...re represented by integers or characters, and each individual ...racterized by its *n* alleles	...e evolution of mutators[135,136] ...acterial speciation in neutral conditions[86]
Network	The individuals are characterized by a graph r... ...enting a gene-regulatory network, a neural network or... ...a logic circuit; there is no explicit DNA level, and mutations c... ...y change the connections or the node numbers in the netw...	...e evolution of network evolvability and ...odularity[122,123,152,153] ...e importance of post-transcriptional regulation[124] ...e relationship of robustness to mutations and to noise[154] ...e evolution of communication, cooperation and ...truism[89,155]
String-of-pearls	The genome is a variable-length string of 'pea... ...different types: phenotype genes, transcription factor g... ...repeats, retrotransposons, binding sites, and so on; eac... ...rl type can exist in a predefined number of variants; gene... ...er, order and regulation can evolve through mutations an... ...gements	...enome and network evolvability[141,143] ...esource processing in ecosystems[85] ...ympatric speciation[156]
Sequence-of-nucleotides	The genome is a variable-length string of chara... ...rs; predefined signal sequences, analogous to promoters, terminators or start–stop codons, are used to detect genes; point mutations, indels and rearrangements can be simulated in a realistic manner	...e evolution of non-coding DNA and gene number[140] • The evolution of the size and topology of gene networks[126,127] • Gene network inference[157,158]

*Many formalisms have been proposed to represent the genome, each with strengths and weaknesses. The appropriate formalism strongly depends on the question of interest. Here, we focus on the approaches that are most directly comparable to *in vivo* microbial evolution experiments (that is, approaches for which the genome comprises several genes).

Molecular realism

Computability

Table 2 | **Genome formalisms in *in silico* experimental evolution**

Formalism*	Description	Questions addressed
Program	The genome is a sequence of instructions in a programing language. The fitness of the program depends on its ability to create copies of itself in the computer's memory and/or to perform specific computations	• The emergence of parasites and hyperparasites[38] • The evolution of robustness, evolvability, complexity and modularity[3,87,142,147,150] • The adaptive radiation of species[84] • The information threshold (the maximum amount of information that can be evolutionarily maintained)[151]
Allelic	The genome is made up of a fixed gene number, *n*; each gene can exist in a finite or infinite number of alleles; alleles are represented by integers or characters, and each individual is characterized by its *n* alleles	• The evolution of mutators[135,136] • Bacterial speciation in neutral conditions[86]
Network	The individuals are characterized by a graph representing a gene-regulatory network, a neural network or even a logic circuit; there is no explicit DNA level, and mutations directly change the connections or the node numbers in the network	• The evolution of network evolvability and modularity[122,123,152,153] • The importance of post-transcriptional regulation[124] • The relationship of robustness to mutations and to noise[154] • The evolution of communication, cooperation and altruism[89,155]
String-of-pearls	The genome is a variable-length string of 'pearls' of different types: phenotype genes, transcription factor genes, repeats, retrotransposons, binding sites, and so on; each pearl type can exist in a predefined number of variants; gene number, order and regulation can evolve through mutations and rearrangements	• Genome and network evolvability[141,143] • Resource processing in ecosystems[85] • Sympatric speciation[156]
Sequence-of-nucleotides	The genome is a variable-length string of characters; predefined signal sequences, analogous to promoters, terminators or start–stop codons, are used to detect genes; point mutations, indels and rearrangements can be simulated in a realistic manner	• The evolution of non-coding DNA and gene number[140] • The evolution of the size and topology of gene networks[126,127] • Gene network inference[157,158]

*Many formalisms have been proposed to represent the genome, each with strengths and weaknesses. The appropriate formalism strongly depends on the question of interest. Here, we focus on the approaches that are most directly comparable to *in vivo* microbial evolution experiments (that is, approaches for which the genome comprises several genes).

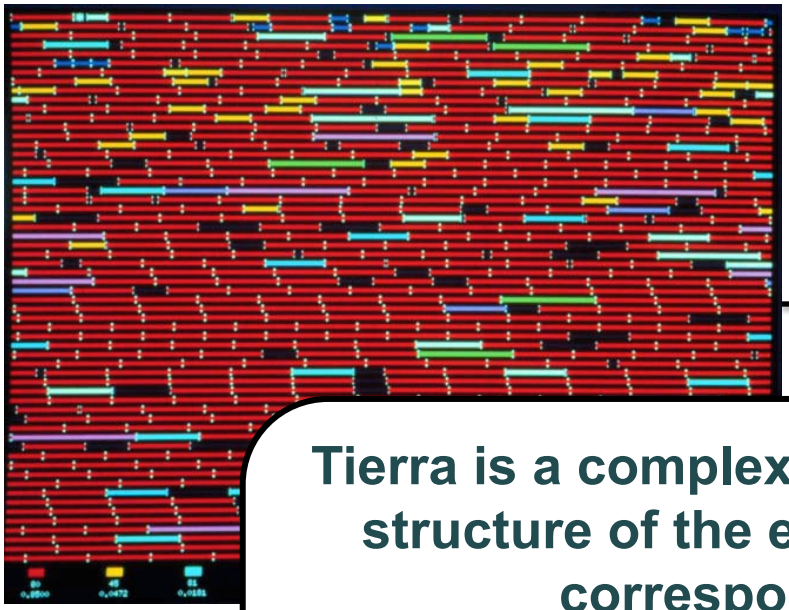
Tierra: the ancestor
(Tom Ray, 1992)

- Tierra is an evolving artificial ecosystem
  - → In biology organisms use energy to organize matter
  - → In Tierra programs use CPU time to organize memory

"*In Tierra, the self-replicating entities are executable machine code programs, which do nothing more than make copies of themselves in the RAM memory of the computer. Thus the machine code becomes an analogue of the nucleic acid based genetic code of organic life"* (T. Ray)

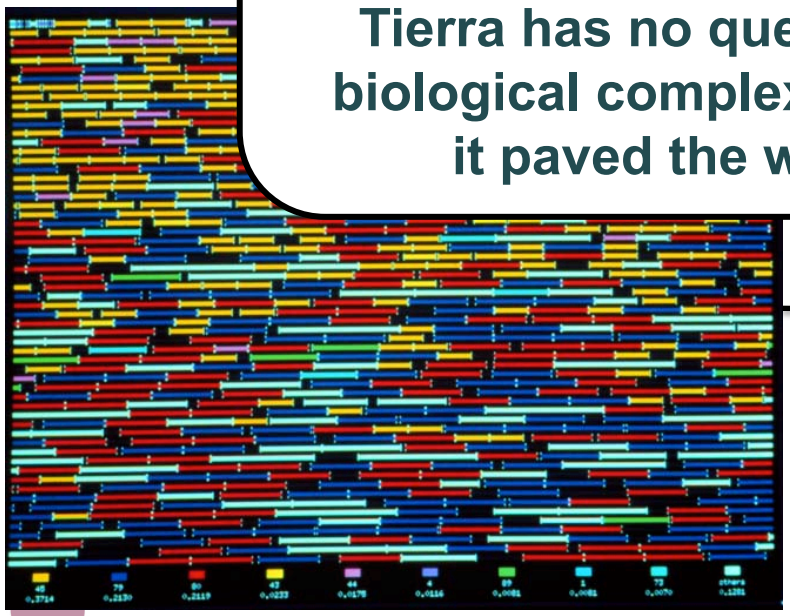
- Tierra enables to study the evolutionary behavior of evolving entities engaged in an "open-ended evolution"
  - No goal but survive and reproduce
  - Starting with a hand-made self-replicating "program-organism" (80 instructions), just let it evolve and look at what happens…

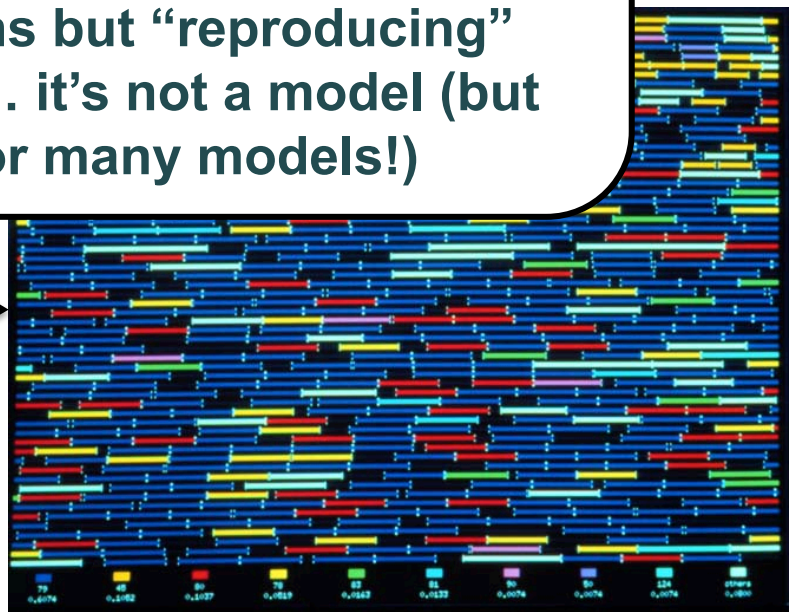
→ http://life.ou.edu/



Tierra is a complex ecological model… but the structure of the evolved organisms has no correspondence in biology

Tierra has no questions but "reproducing" biological complexity … it's not a model (but it paved the way for many models!)

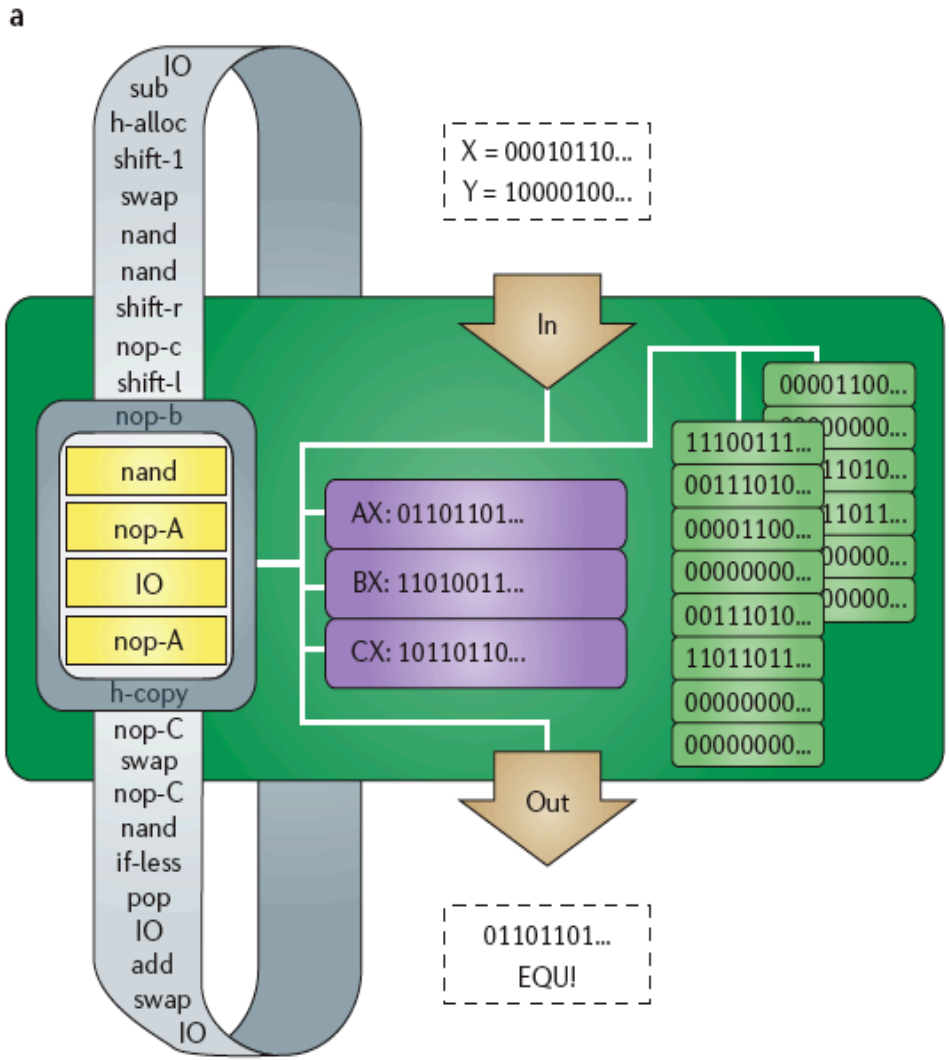




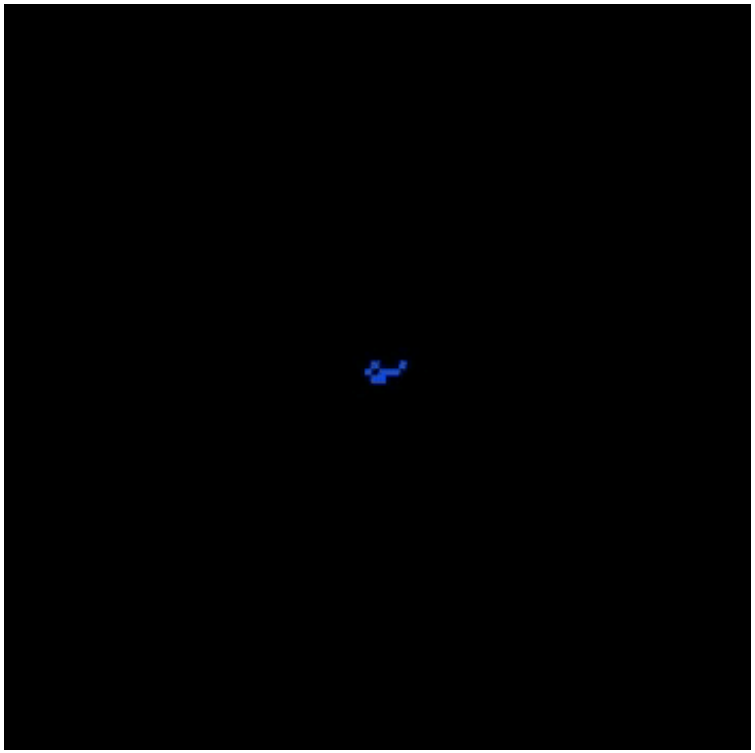
Avida (adami, 1998)

- Compared with Tierra, Avida is not a "better" model; it is a new way to use models in alife
  - Experimental method, biological questions ("origin of complexity")
  - Permanent interactions with biologists…
- Avida also uses a simpler artificial chemistry than Tierra
  - Each "avidian" contains its own CPU (no interactions during code execution) → parasitism/symbiosis are no longer possible
  - Avidians are immerged in a 2D space
  - The evolution is no more open-ended, results are easier to analyze!
  - Better trade-of between simplicity and complexity of the model?
- Many results in biology
  - See e.g., C. Adami, T. Collier, S. F. Elena, C. Ofria, C. Wilke, R. Lenski, D. Misevic…

“Avidians”



(C. Adami, *Nat. Rev. Genet.*, 2006)

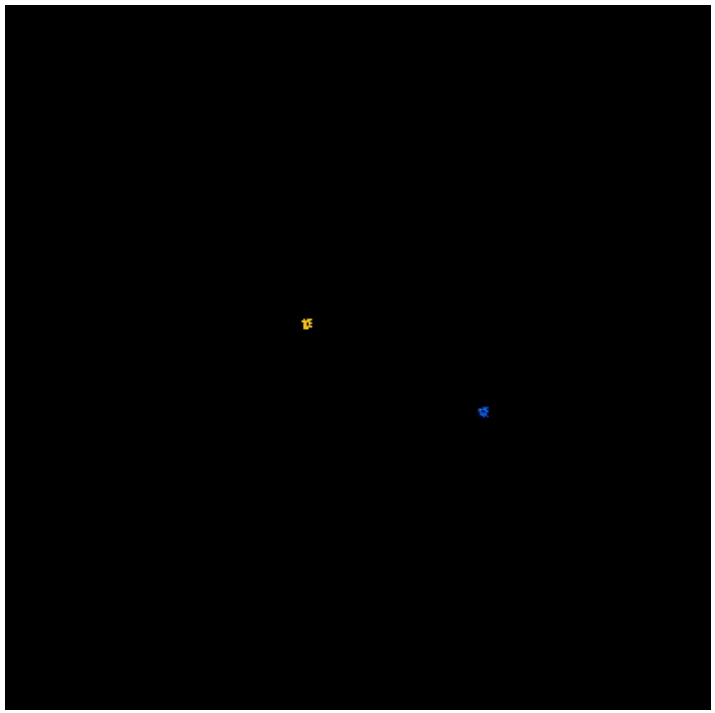


All simulation start from a man-made self-replicating organism (like in Tierra)

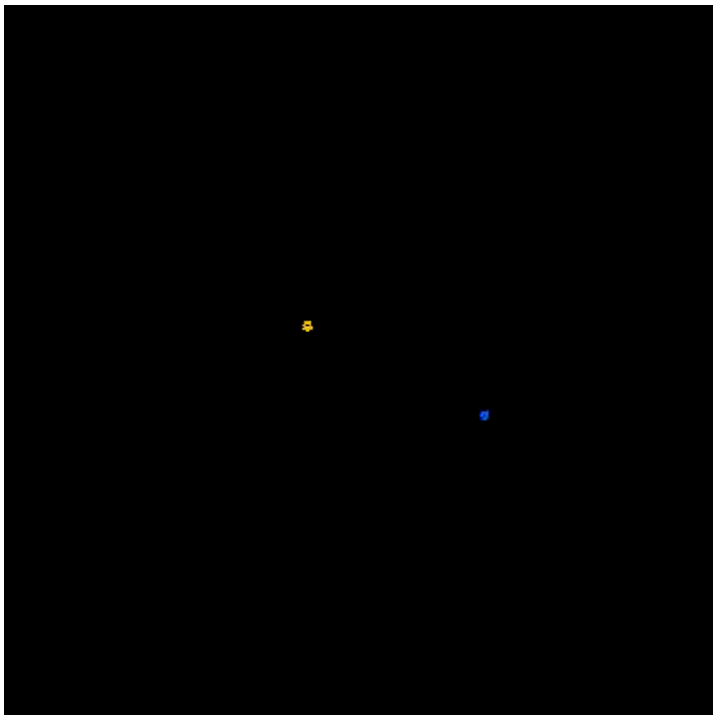
Experimenting with avida

- Competition assay: Two different organisms in a same (virtual) “petri dish”
  - Yellow organism: good but not robust
  - Blue organism = not so good but robust

Mutation rate: 0.5 per replic.



Mutation rate: 1.5 per replic.



Evolution of digital organisms at high mutation rates leads to survival of the flattest

Claus O. Wilke*, Jia Lan Wang*, Charles Ofria†, Richard E. Lenski† & Christoph Adami*‡

* *Digital Life Laboratory, Mail Code 136-93, Cali*
Pasadena, California 91125, USA
† *Center for Biological Modeling, Michigan State*
Michigan 48824, USA
‡ *Jet Propulsion Laboratory, Mail Code 126-347,*
Technology, Pasadena, California 91109, USA

Darwinian evolution favours genotype rates, a process called 'survival of the fi the replication rate of each individual ge predict the eventual survivor, even in According to quasi-species theory, selec genotypes, interconnected by mutation tion rate is highest[1–5]. Here we confir digital organisms that self-replicate, mu

NATURE | VOL 412 | 19 JULY 2001 | www.nature.com

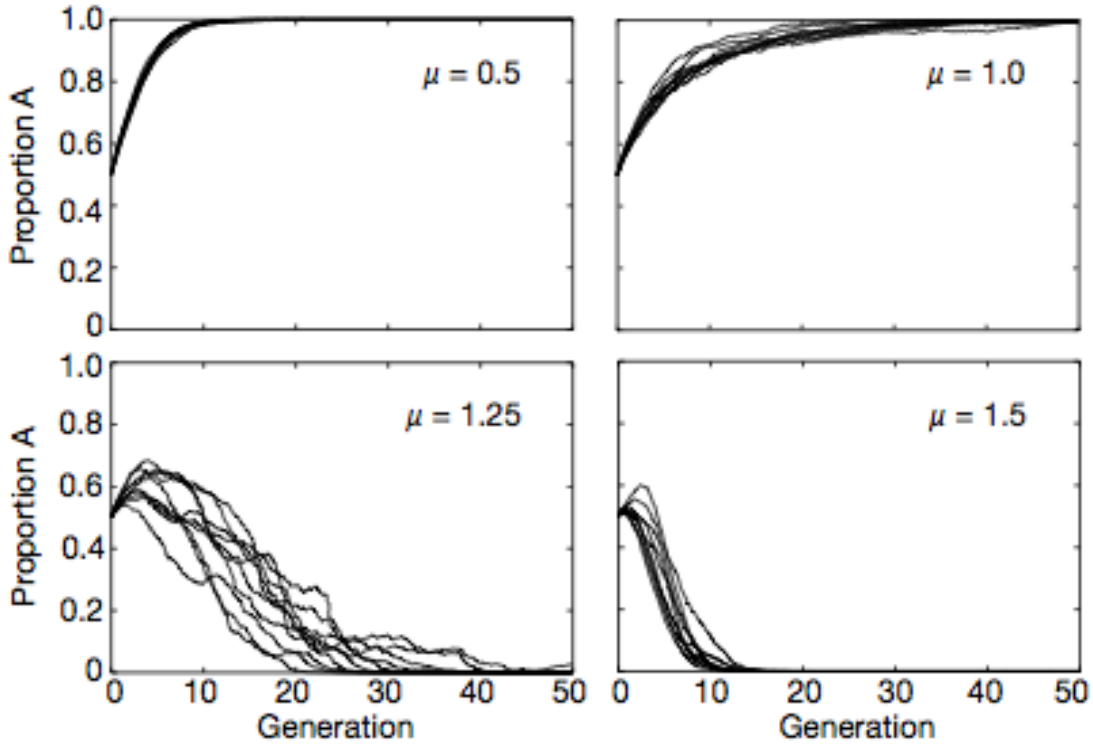
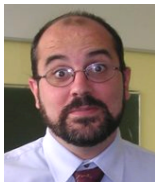


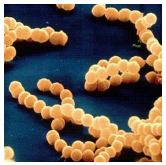
Figure 1 Competitions for one pair of organisms at four different mutation rates. Organism A replicates 1.96 times faster than B. μ, Genomic mutation rate.

But many open questions cannot be addressed with Avida



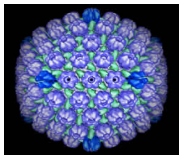
Homo sapiens

~3 billions bp
~25 000 genes



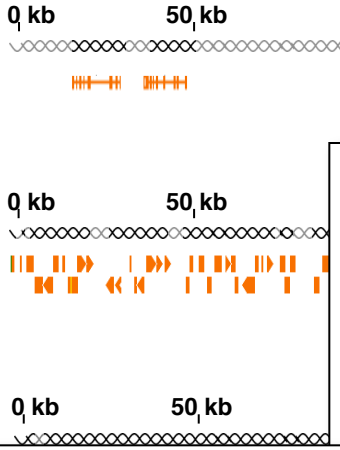
Neisseria meningitidis

~2 millions bp
~2 000 genes



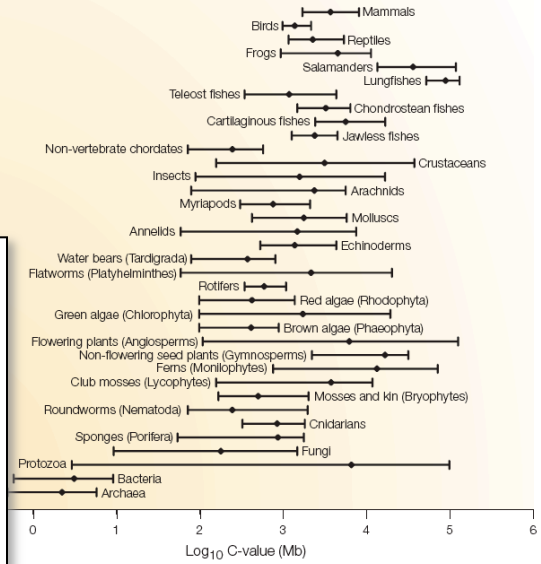
Herpes HS

~150 000 bp
~100 genes

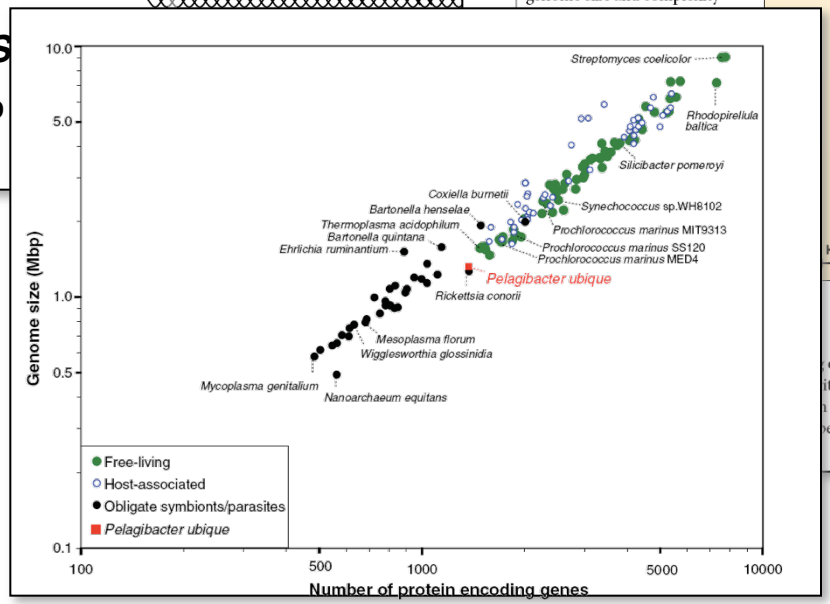


Box 1 | **Extensive variation in genome size within and among the main groups of life**

Ever since the first general surveys of nuclear DNA content were carried out in the early 1950s it has been apparent that eukaryotic genome sizes vary enormously and that this is unrelated to intuitive ideas of morphological complexity[2]. This discrepancy between genome size and complexity



organisms, and are loosely arranged according to common ideas of ity between this parameter and genome size. Some commonly cited extreme omitted, as there is considerable uncertainty about the accuracy of these ecies involved[10,87].



The "sequence of nucleotides" formalism

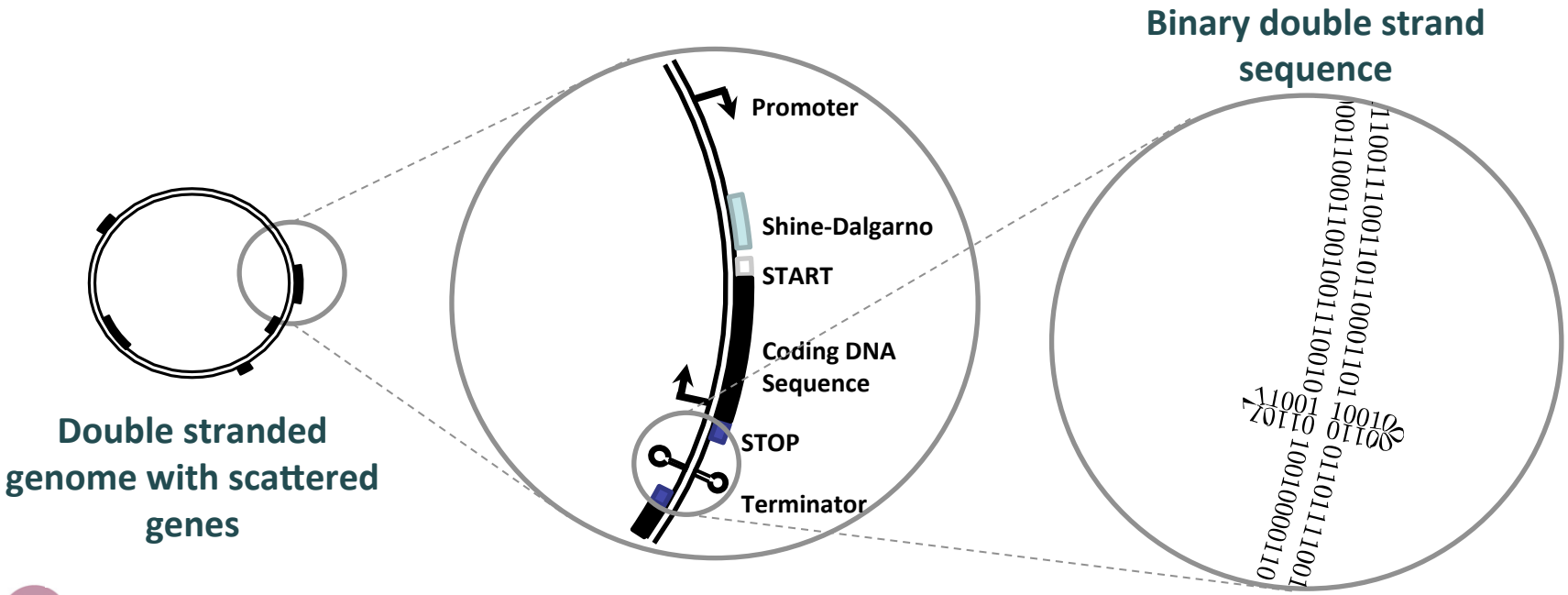
Table 2 | **Genome formalisms in *in silico* experimental evolution**

Formalism*	Description	Questions addressed
Program	The genome is a sequence of instructions in a programing language. The fitness of the program depends on its ability to create copies of itself in the computer's memory and/or to perform specific computations	• The emergence of parasites and hyperparasites[38] • The evolution of robustness, evolvability, complexity and modularity[3,87,142,147,150] • The adaptive radiation of species[84] • The information threshold (the maximum amount of information that can be evolutionarily maintained)[151]
Allelic	The genome is made up of a fixed gene number, *n*; each gene can exist in a finite or infinite number of alleles; alleles are represented by integers or characters, and each individual is characterized by its *n* alleles	• The evolution of mutators[135,136] • Bacterial speciation in neutral conditions[86]
Network	The individuals are characterized by a graph representing a gene-regulatory network, a neural network or even a logic circuit; there is no explicit DNA level, and mutations directly change the connections or the node numbers in the network	• The evolution of network evolvability and modularity[122,123,152,153] • The importance of post-transcriptional regulation[124] • The relationship of robustness to mutations and to noise[154] • The evolution of communication, cooperation and altruism[89,155]
String-of-pearls	The genome is a variable-length string of 'pearls' of different types: phenotype genes, transcription factor genes, repeats, retrotransposons, binding sites, and so on; each pearl type can exist in a predefined number of variants; gene number, order and regulation can evolve through mutations and rearrangements	• Genome and network evolvability[141,143] • Resource processing in ecosystems[85] • Sympatric speciation[156]
Sequence-of-nucleotides	The genome is a variable-length string of characters; predefined signal sequences, analogous to promoters, terminators or start–stop codons, are used to detect genes; point mutations, indels and rearrangements can be simulated in a realistic manner	• The evolution of non-coding DNA and gene number[140] • The evolution of the size and topology of gene networks[126,127] • Gene network inference[157,158]

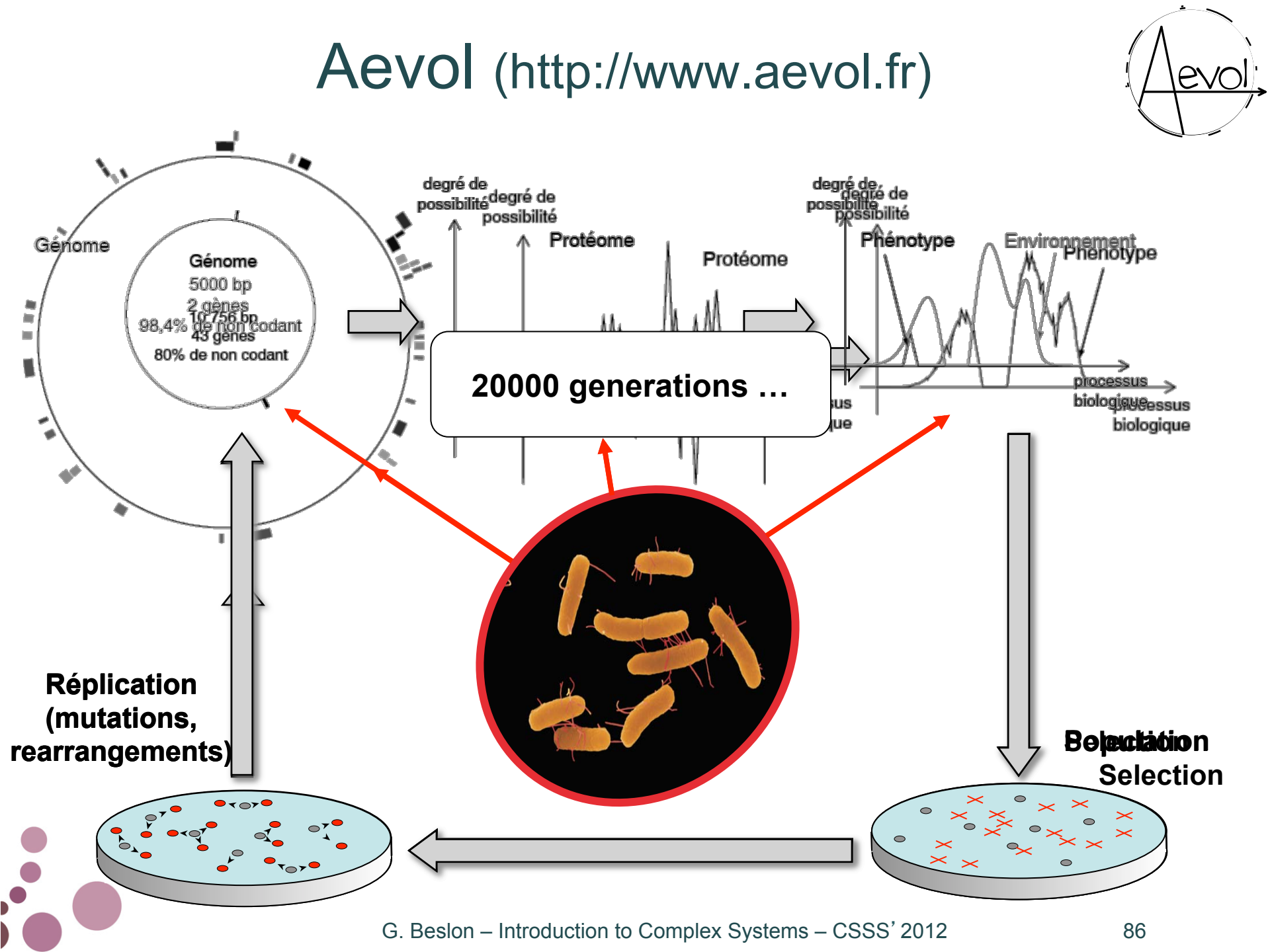
*Many formalisms have been proposed to represent the genome, each with strengths and weaknesses. The appropriate formalism strongly depends on the question of interest. Here, we focus on the approaches that are most directly comparable to *in vivo* microbial evolution experiments (that is, approaches for which the genome comprises several genes).

The “sequence of nucleotides” formalism

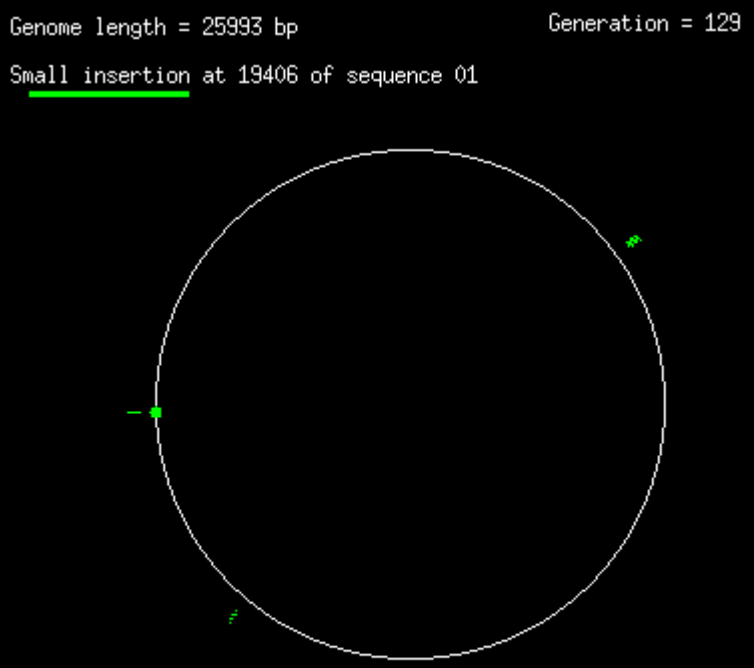
- Organisms own a genome made of a string of nucleotides (often 0/1, rarely ACTG)
- The genome is translated into a phenotype thanks to a process inspired from the “central dogma”
- Mutations can be accurately modeled and their effect on the emergence of complex structures can be studied in the model



Aevol (http://www.aevol.fr)



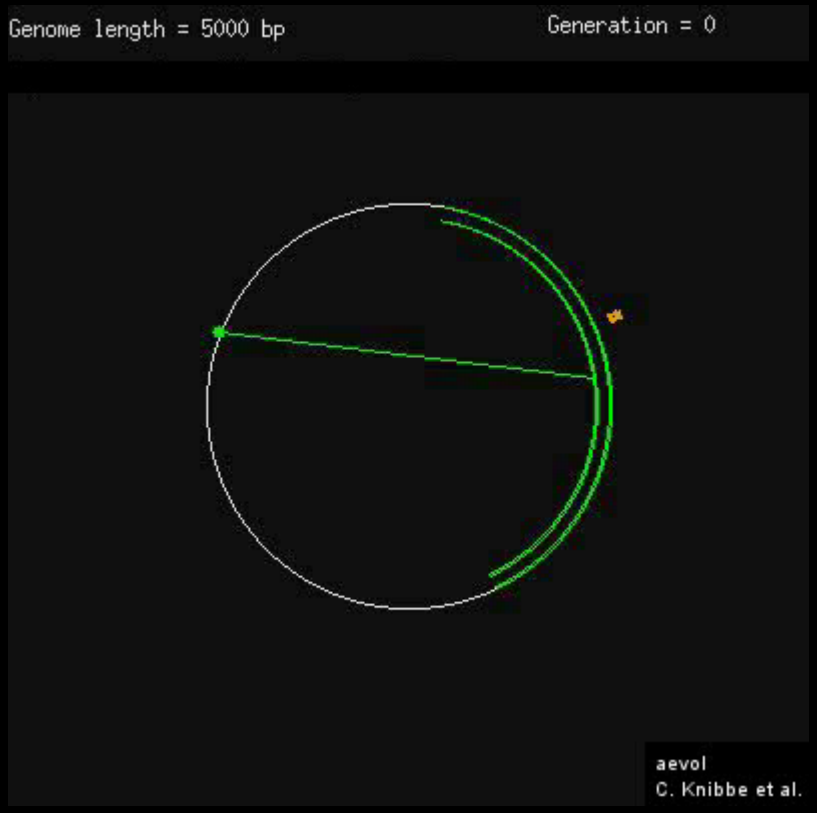
aevol: looking at evolution in action (« winning » lineage)



In-silico experimental evolution

- Experimental study of the effect of mutation/rearrangement rates (u) and selection strength (k) on genome structure
  - Mutation rate u:
    - Six mutation rates from $u = 5.10^{-6}$ to $u = 2.10^{-4}$ per bp
    - Same mutation rates for point mutations and rearrangements
  - Selection:
    - Two selection modes (fitness proportional or rank-based)
    - Different selection strength (here $k = 250$ or $k = 1000$)
  - Experimental evolution during 20000 generations
    - Populations: 1000 individuals
    - Steady environment
  - Three repetitions per couple (u,k) per mutation mode
    - More than 100 simulations
    - It's *really* an experimental approach …

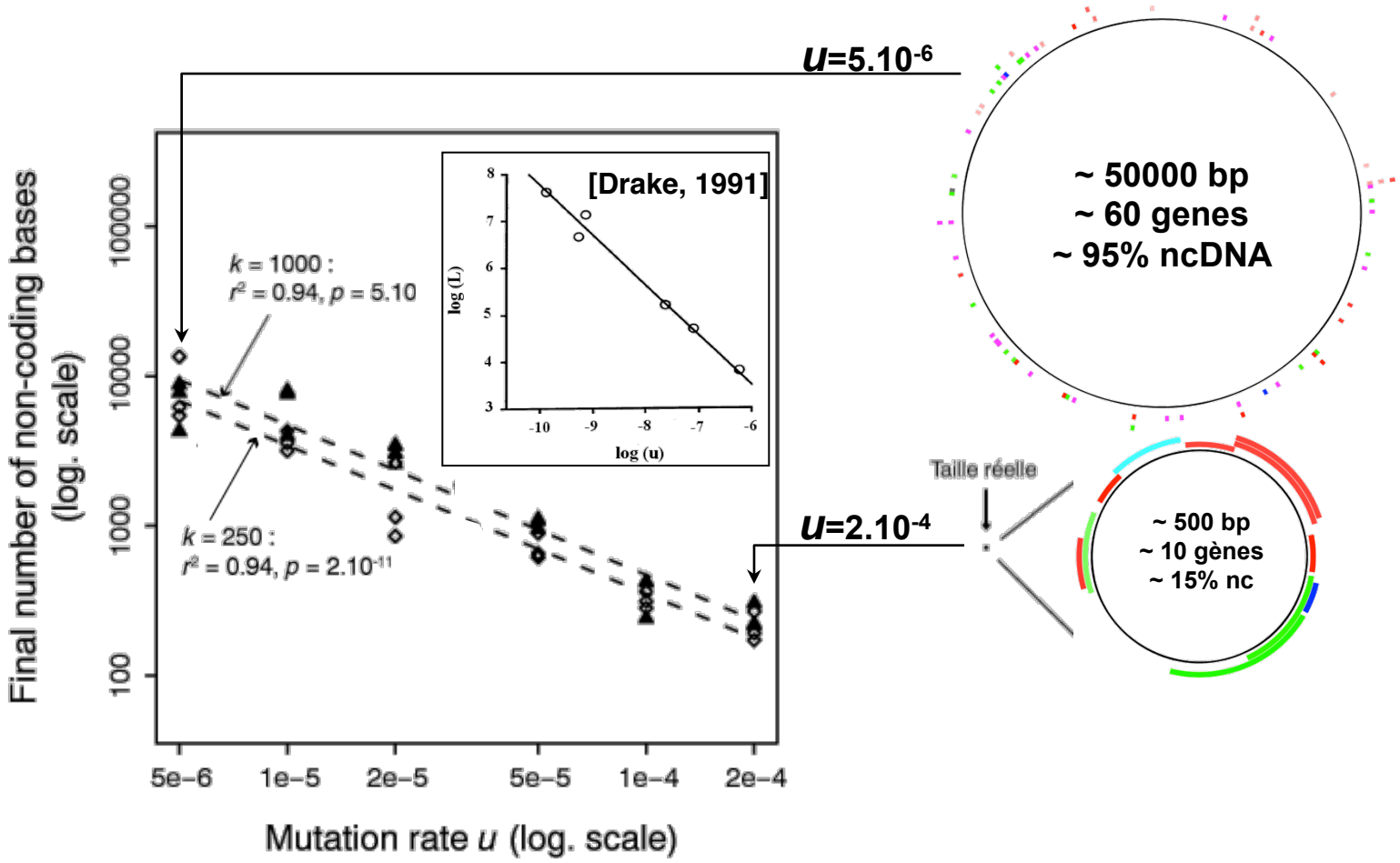
Ævol: The movie (II) ...



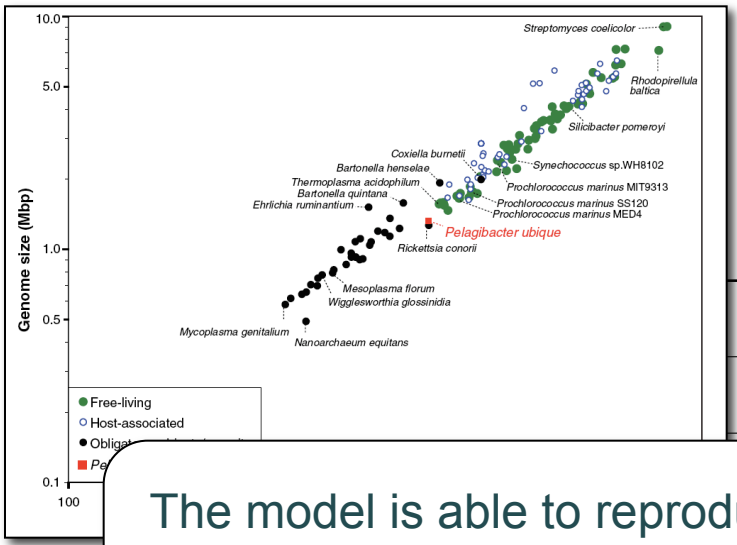
High mutation rates : 2.10^{-4} / pb

Low mutation rates : 5.10^{-6} / pb

In silico experimental evolution



Yet another model explaining everything ;)

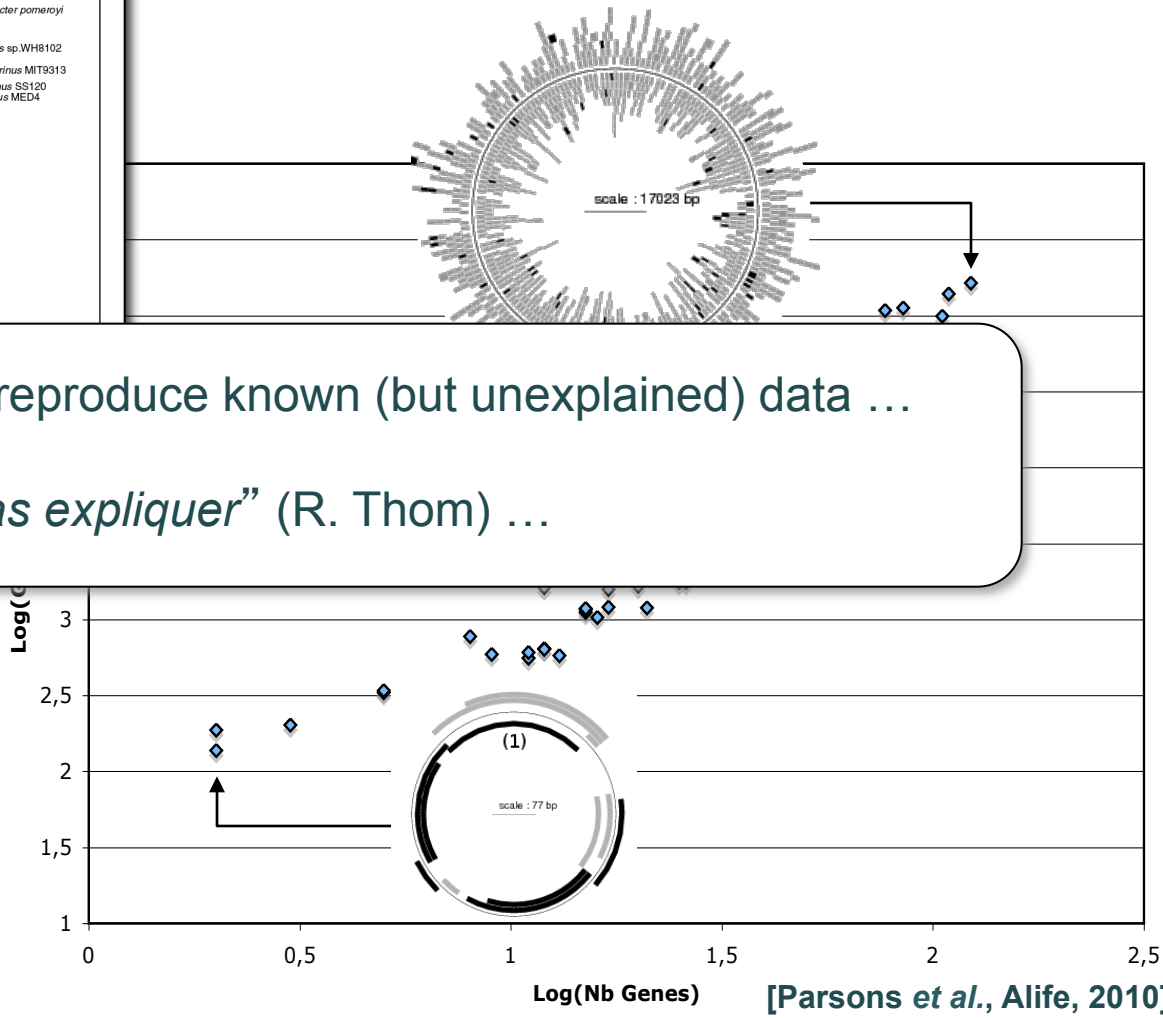


[Giova

The model is able to reproduce known (but unexplained) data …

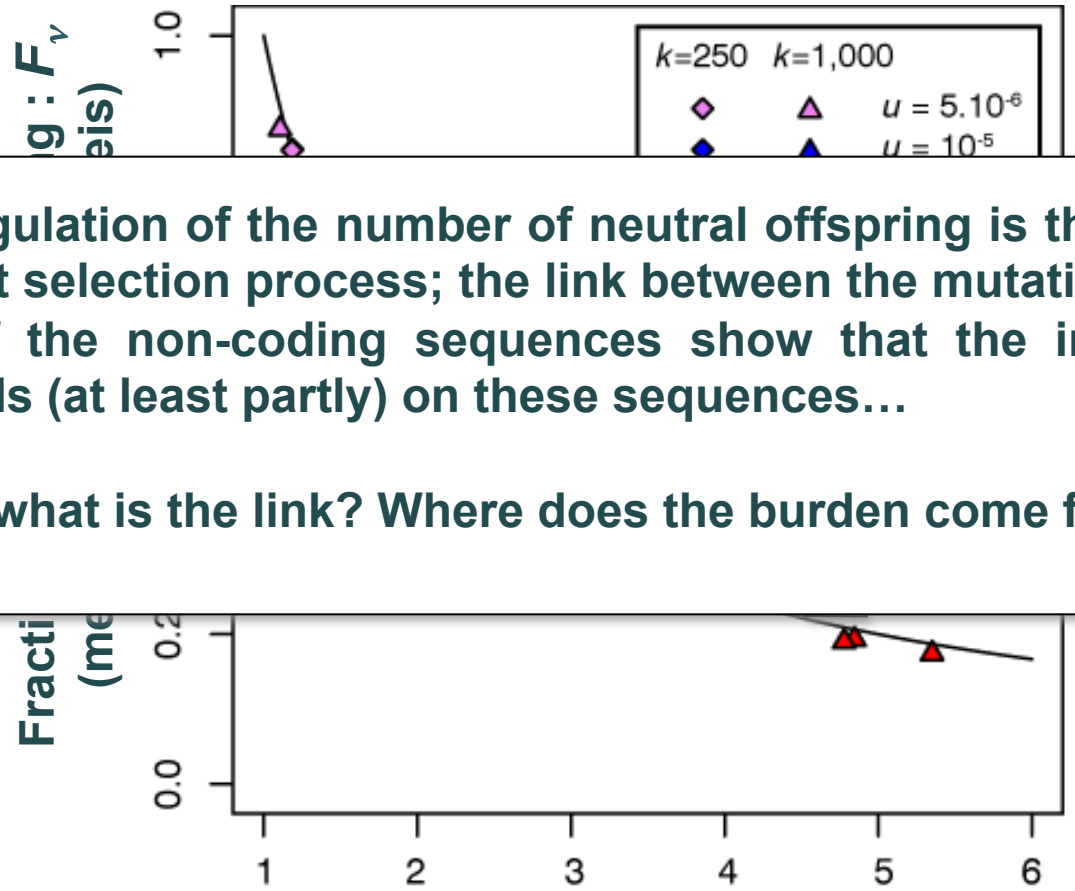
But "*Prédire n'est pas expliquer*" (R. Thom) …

- The m
  observed regularities at the molecular (genome) level …
- Looking at the RNA structures, it predicts that operons are more numerous and RNAs longer in small genomes while ncRNA are more numerous in long genomes…



[Parsons *et al.*, Alife, 2010]

Measure of best individual reproduction neutrality (at t = 20000)



The regulation of the number of neutral offspring is the hallmark of an indirect selection process; the link between the mutation rate *u* and the size of the non-coding sequences show that the indirect selection depends (at least partly) on these sequences…

… But what is the link? Where does the burden come from?

Modeling the model …

- **Mathematical model of reproduction**
  - **The math model is "true" for aevol *AND* for the "real world"…**
- **E_1 : Probability of neutral reproduction as a function of**

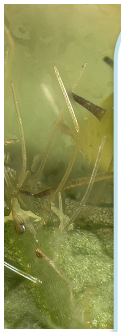
If: (i) genomes undergo large duplications and deletions, (ii) the number and the average size of these events increase with genome size, **Then:** the mutational variability of a lineage depends on the amount of non-coding DNA (it is mutagenic for the genes it surrounds).

Thus the indirect selection for an appropriate level of variability actually selects for a specific amount of non-coding DNA

[Knibbe et al., Mol. Biol. Evol., 2007]

$$\begin{cases} \tilde{\nu}_{\text{ponct}} = \tilde{\nu}_{\text{ins}} = \tilde{\nu}_{\text{del}} = 1 - \frac{l}{L} \\ \tilde{\nu}_{\text{inv}} = \left(1 - \frac{l}{L}\right)^2 \\ \tilde{\nu}_{\text{transloc}} = \left(1 - \frac{l}{L}\right)^3 \end{cases} \qquad \begin{cases} \tilde{\nu}_{\text{gdel}} = \frac{1}{2L^2} \sum_{j=1} \lambda_j \left(\lambda_j + 1\right) \\ \tilde{\nu}_{\text{dup}} = \frac{1}{2L^2} \left(1 - \frac{l}{L}\right) \sum_{j=1}^{N_G} \lambda_j \left(\lambda_j + 1\right) \end{cases}$$

What about gene networks?



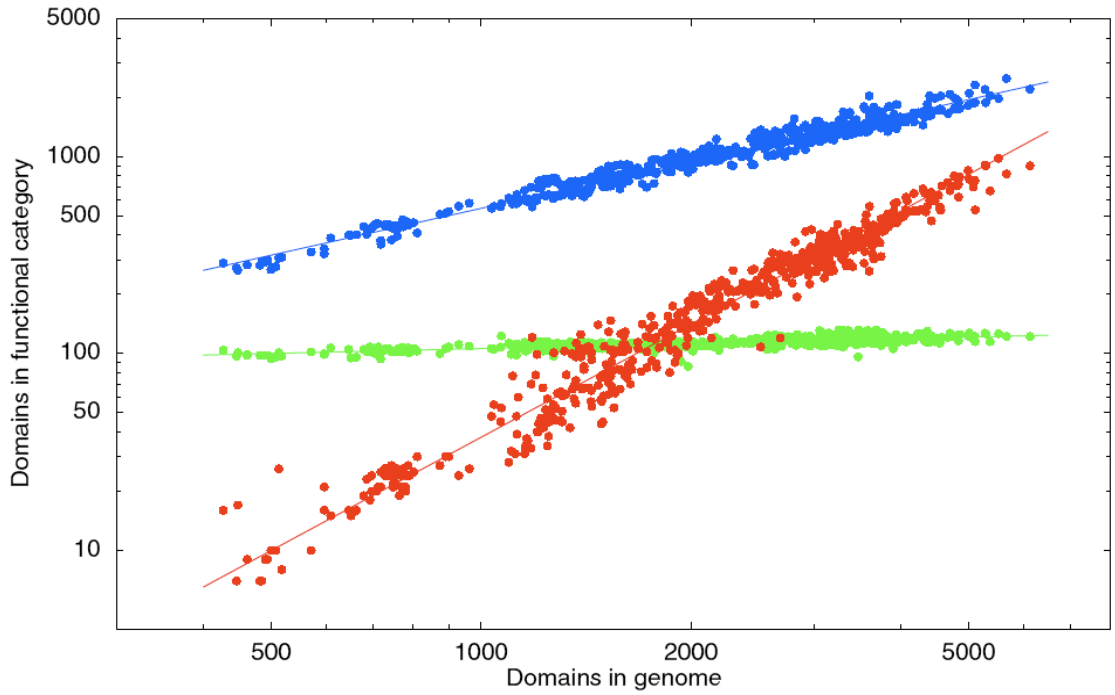
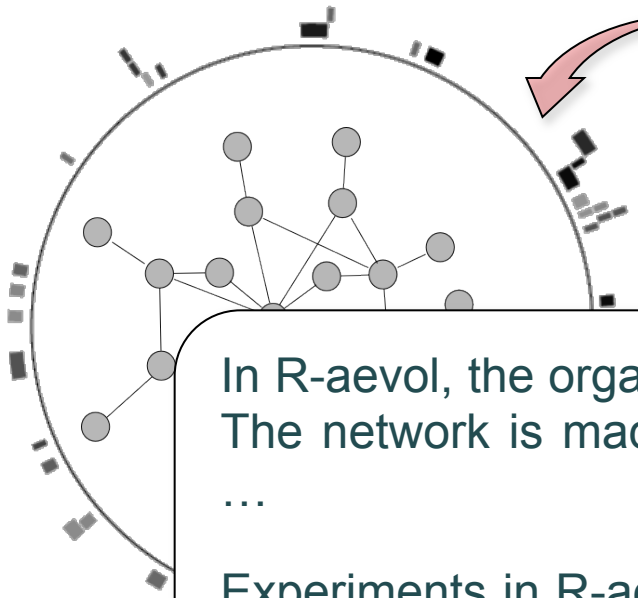
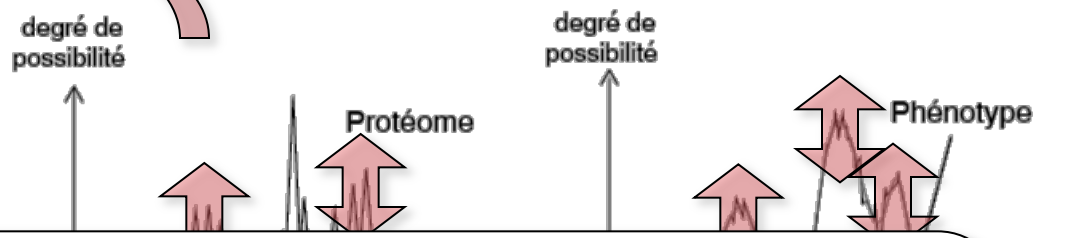


Figure 1
Scaling laws. The number of protein-domains associated with functional categories 'translation' (green), metabolic process' (blue), and 'regulation of transcription' (red) as a function of the total number of domains in the genome for which a functional annotation is available. Each dot corresponds to a fully-sequenced microbial genome, with the total number of domains on the horizontal axis and the number of domains in a particular functional category on the vertical axis. Both axes are shown on a logarithmic scale. The straight lines show power-law fits.

Molina, N., and van Nimwegen, E., The evolution of domain-content in bacterial genomes. Biology Direct (2008) vol. 3 pp. 51

R-aevol: regulation in aevol



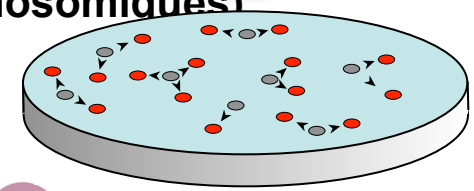


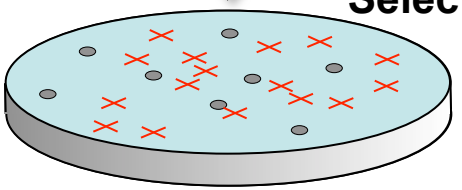
In R-aevol, the organisms own a genome and a regulation network. The network is made of metabolic genes and transcription factors …

Experiments in R-aevol: Does the network structure depend on the evolutionary conditions? (same experiments as in aevol)

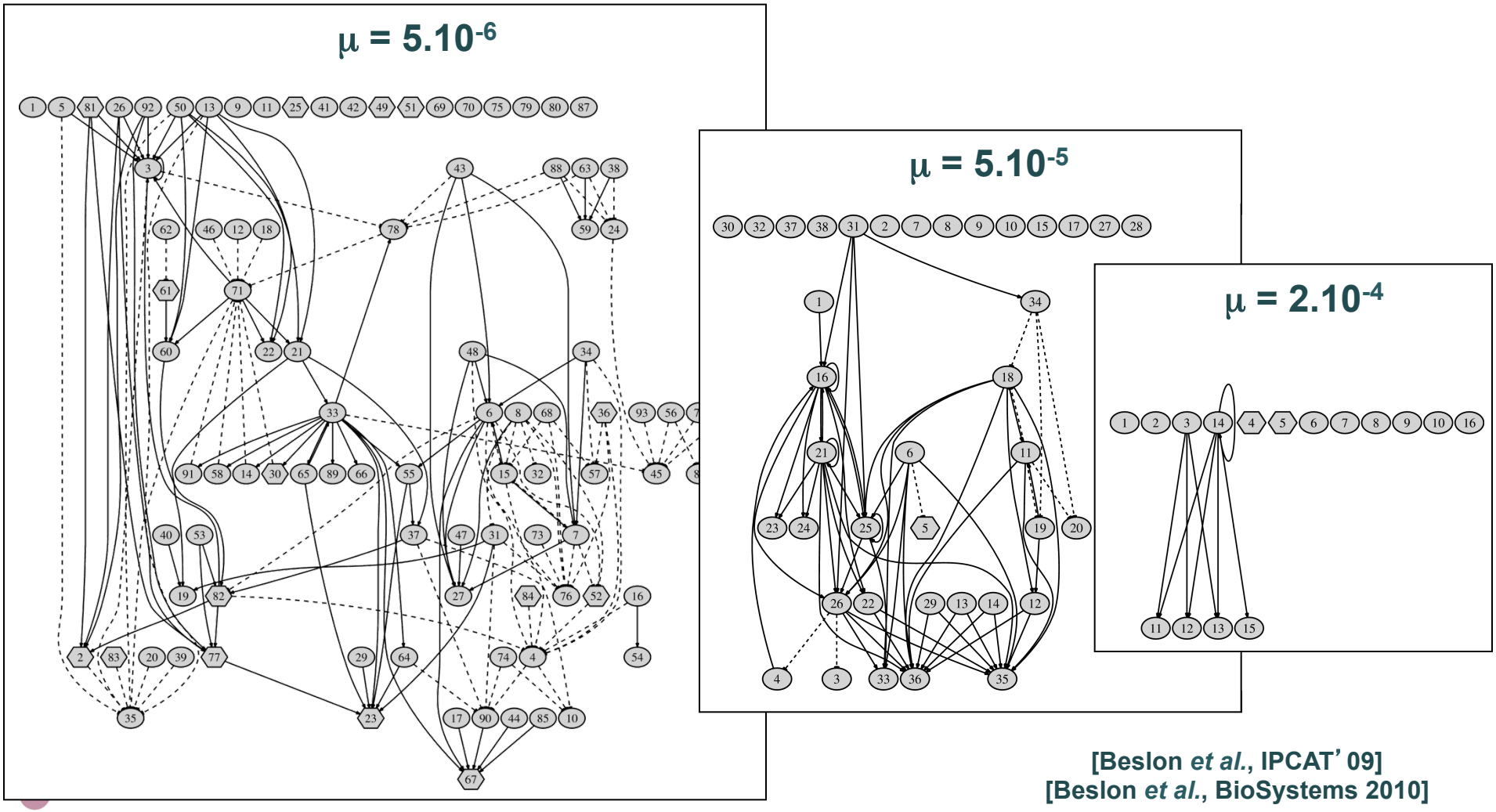
Réplication (mutations, indels réarrangements chromosomiques)

Population Sélection





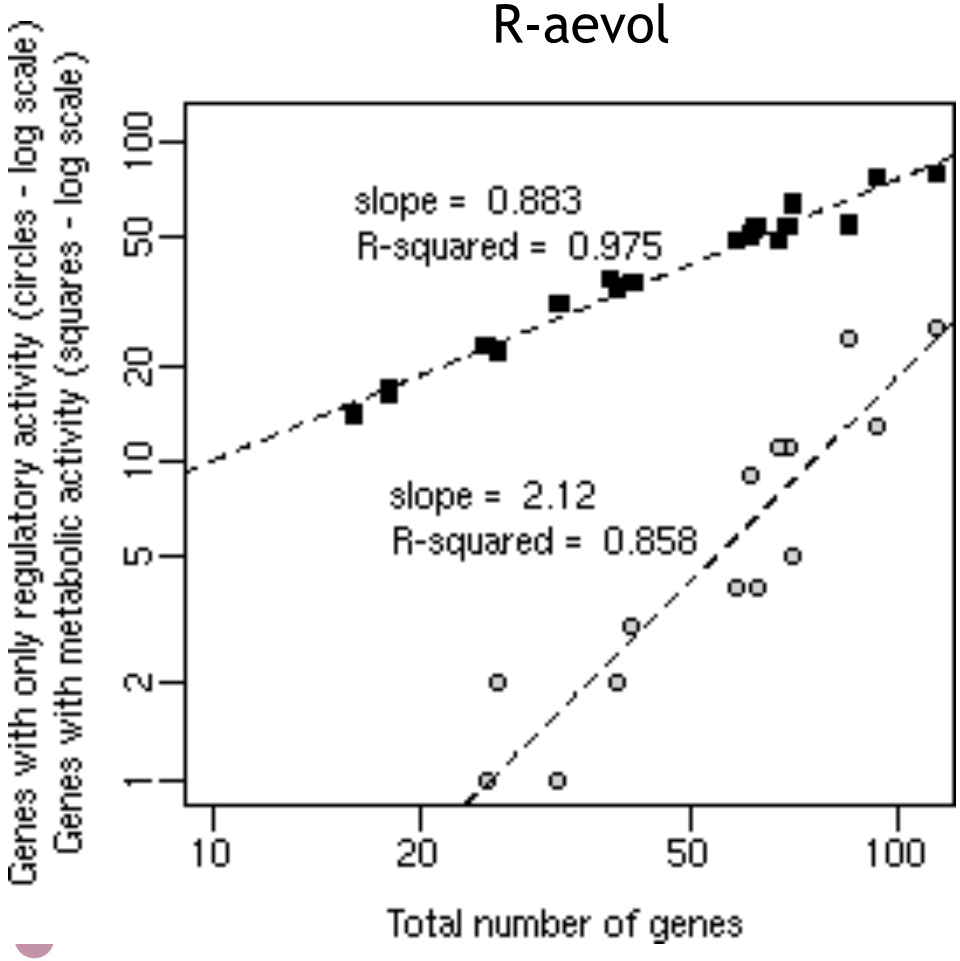
Impact of mutation rates on transcriptomic structures



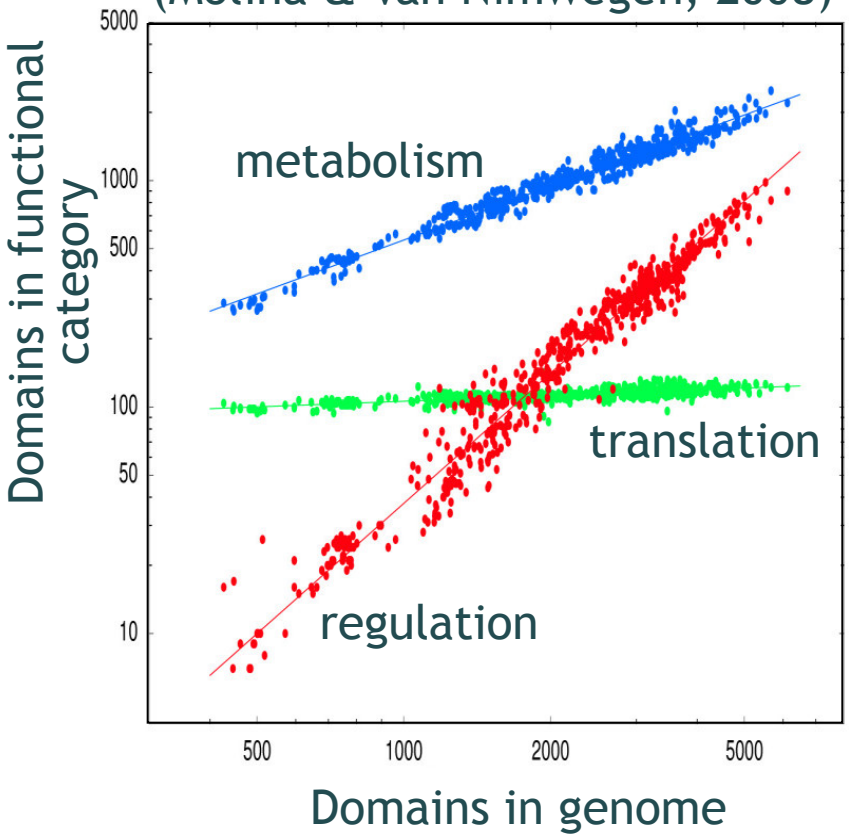
[Beslon *et al.*, IPCAT' 09]
[Beslon *et al.*, BioSystems 2010]

Emergence of scaling laws

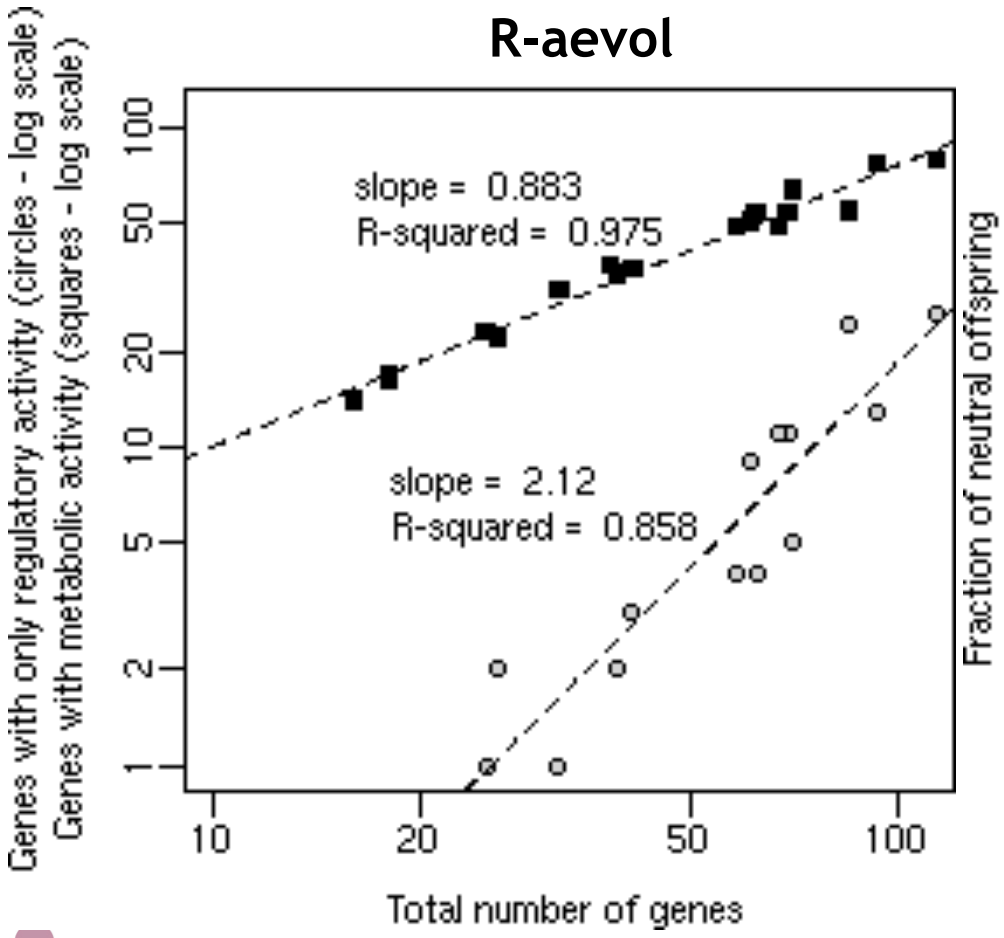
R-aevol

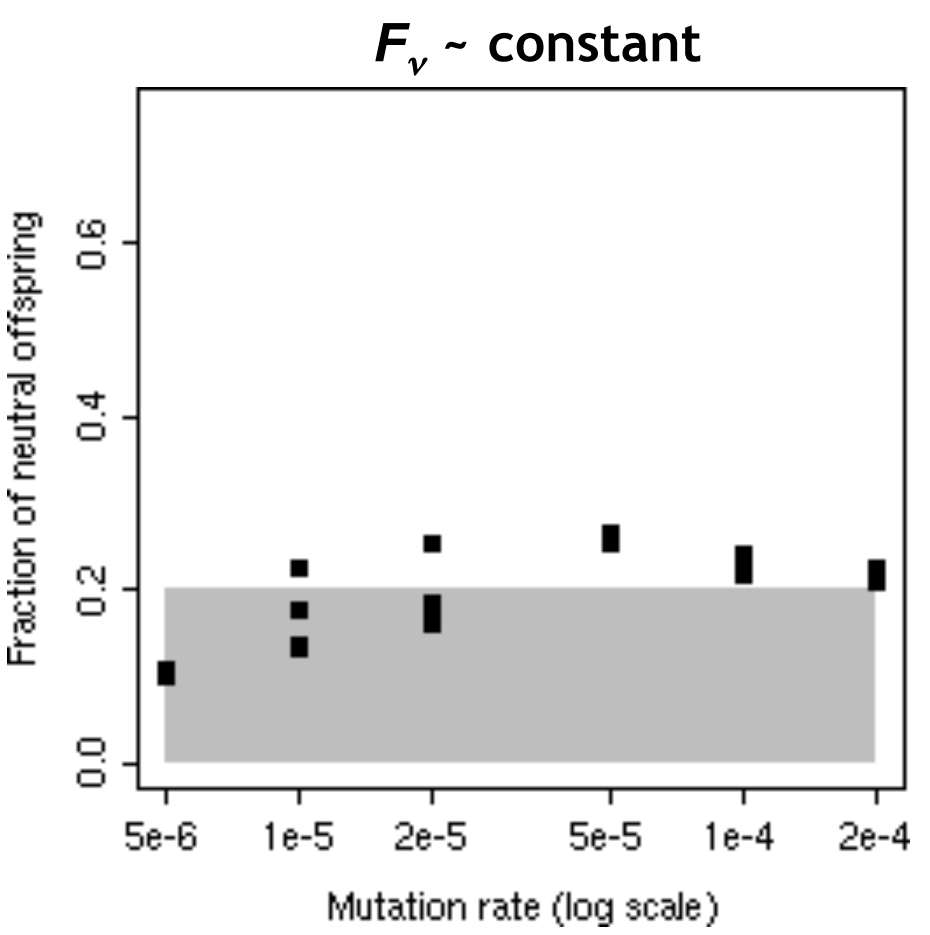


Biological data
(Molina & Van Nimwegen, 2008)



Side effect of the selection for robustness?

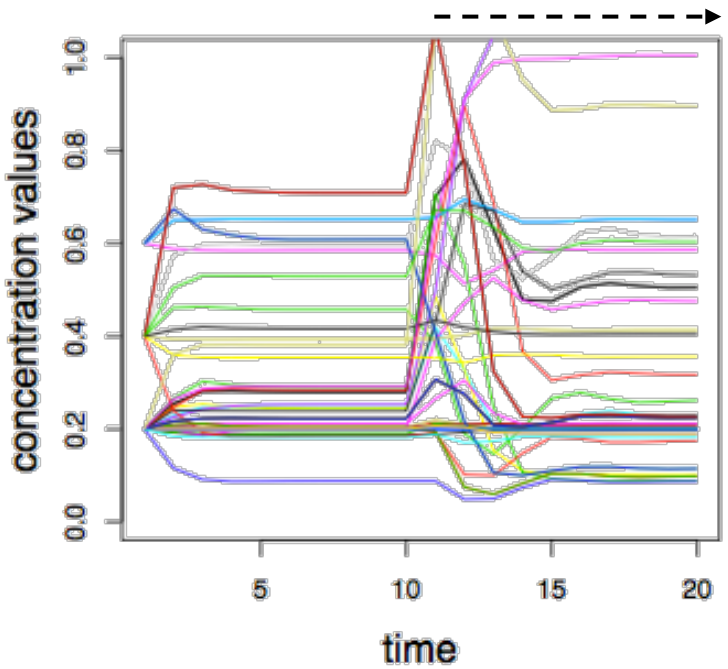




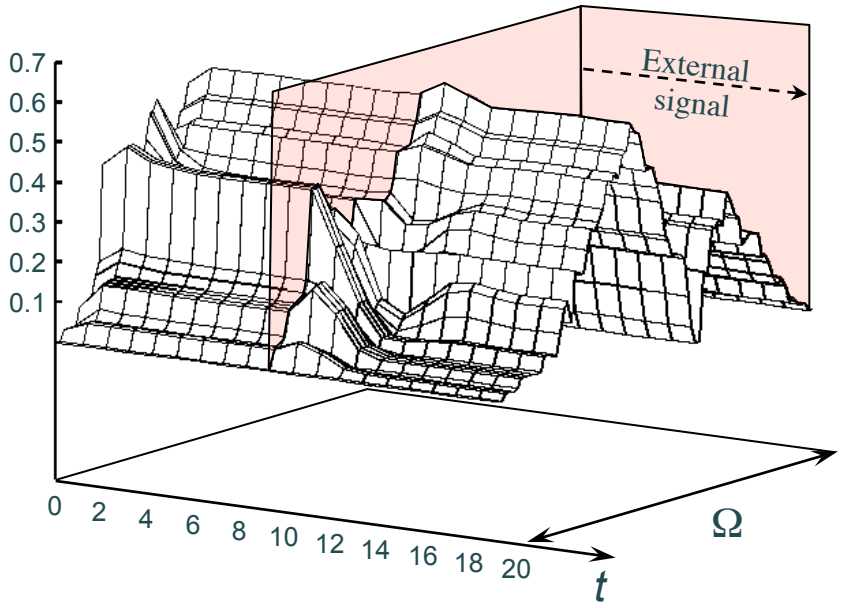
“Two-states” environments

- Organisms live for 20 time steps ; at t = 10 a signal is sent to the “cells” which must react by changing their phenotypes…

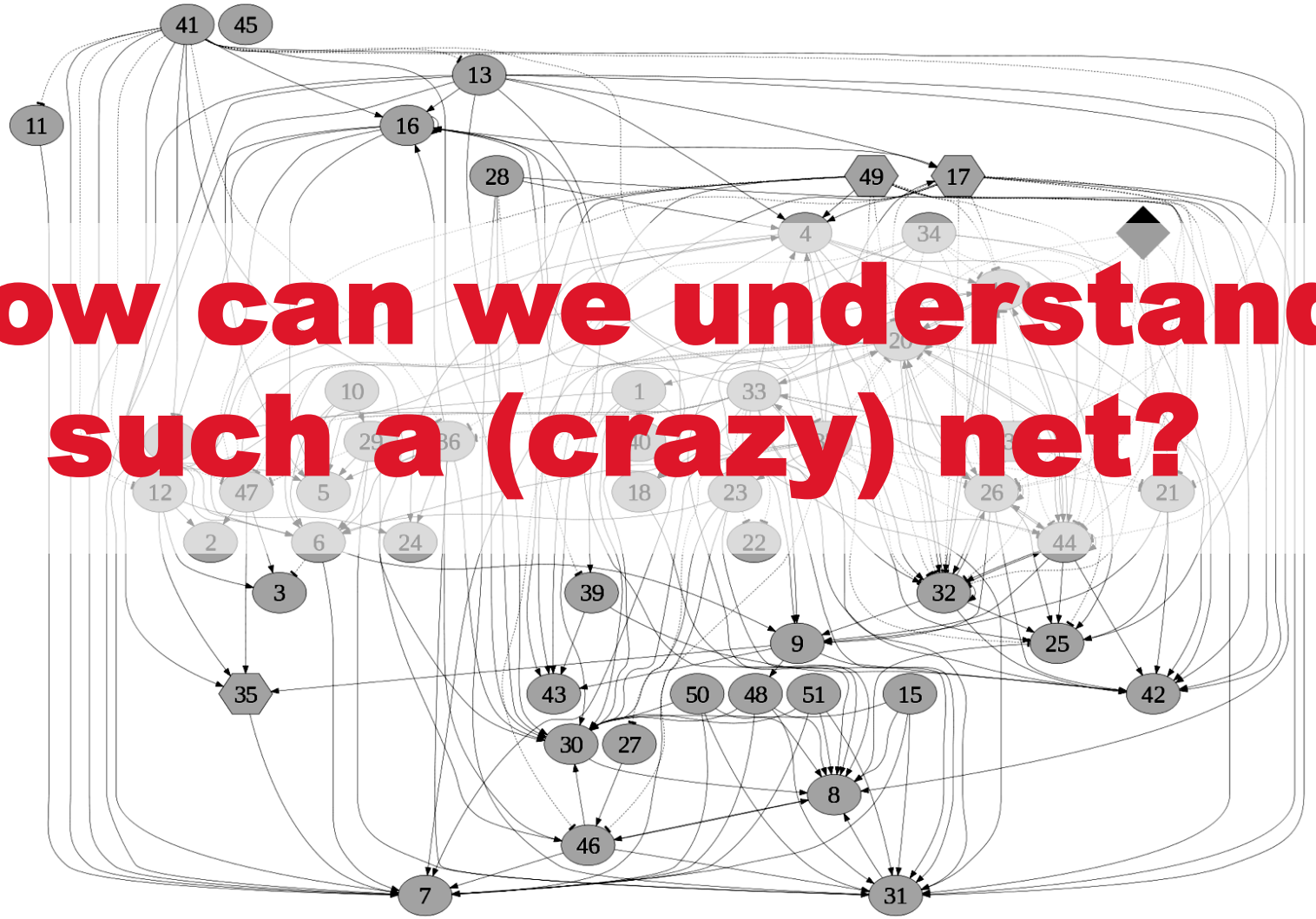
Protein concentrations over time



Phenotype over time

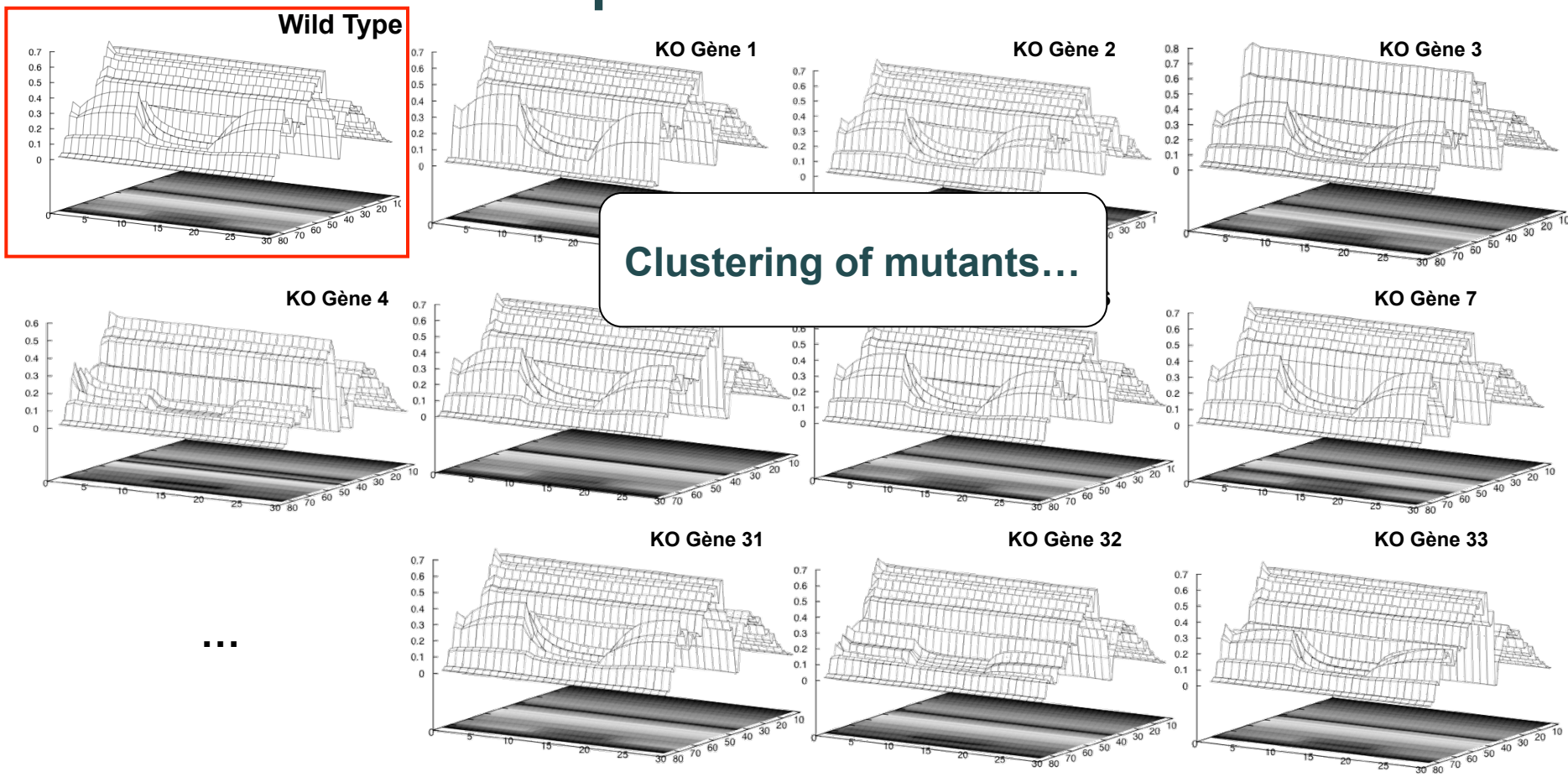


Evolved network after 15000 generations



How can we understand such a (crazy) net?

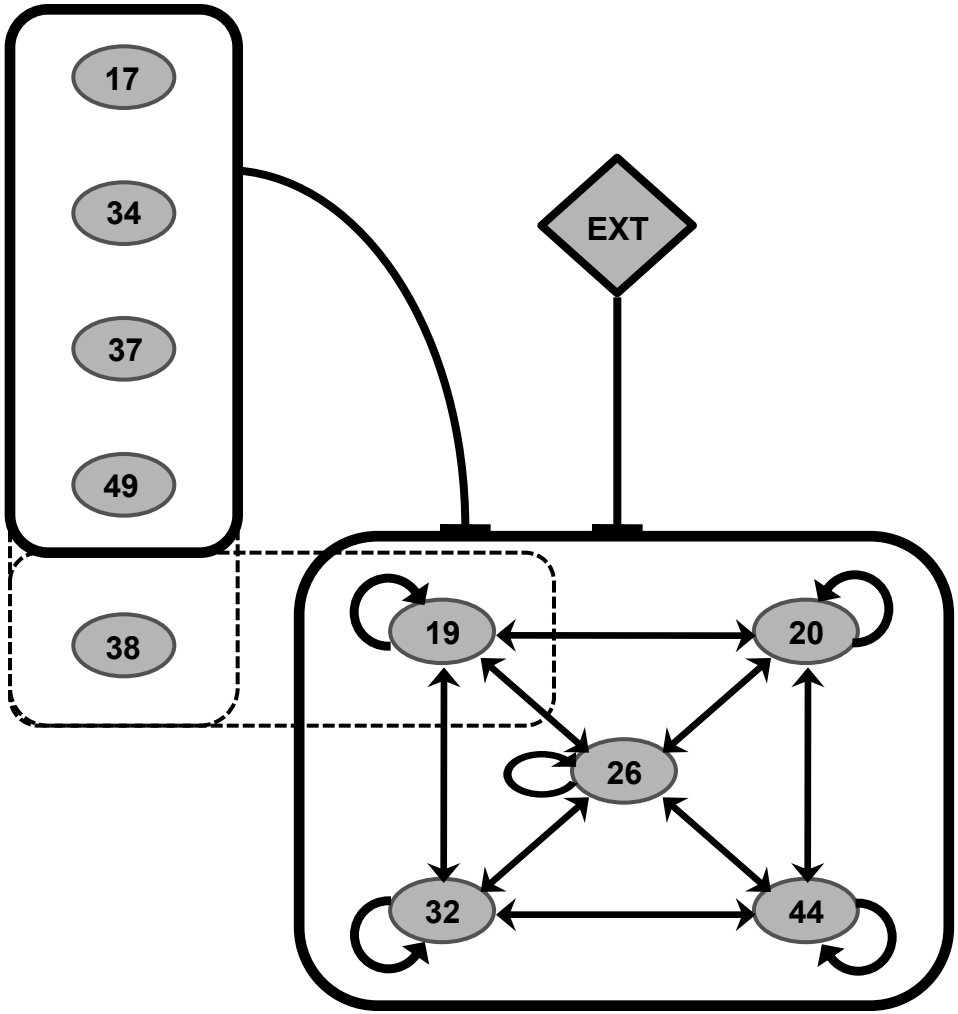
Systematic Knock-Out experiments



Clustering of mutants…

[Beslon et al., IDAj, 2010]

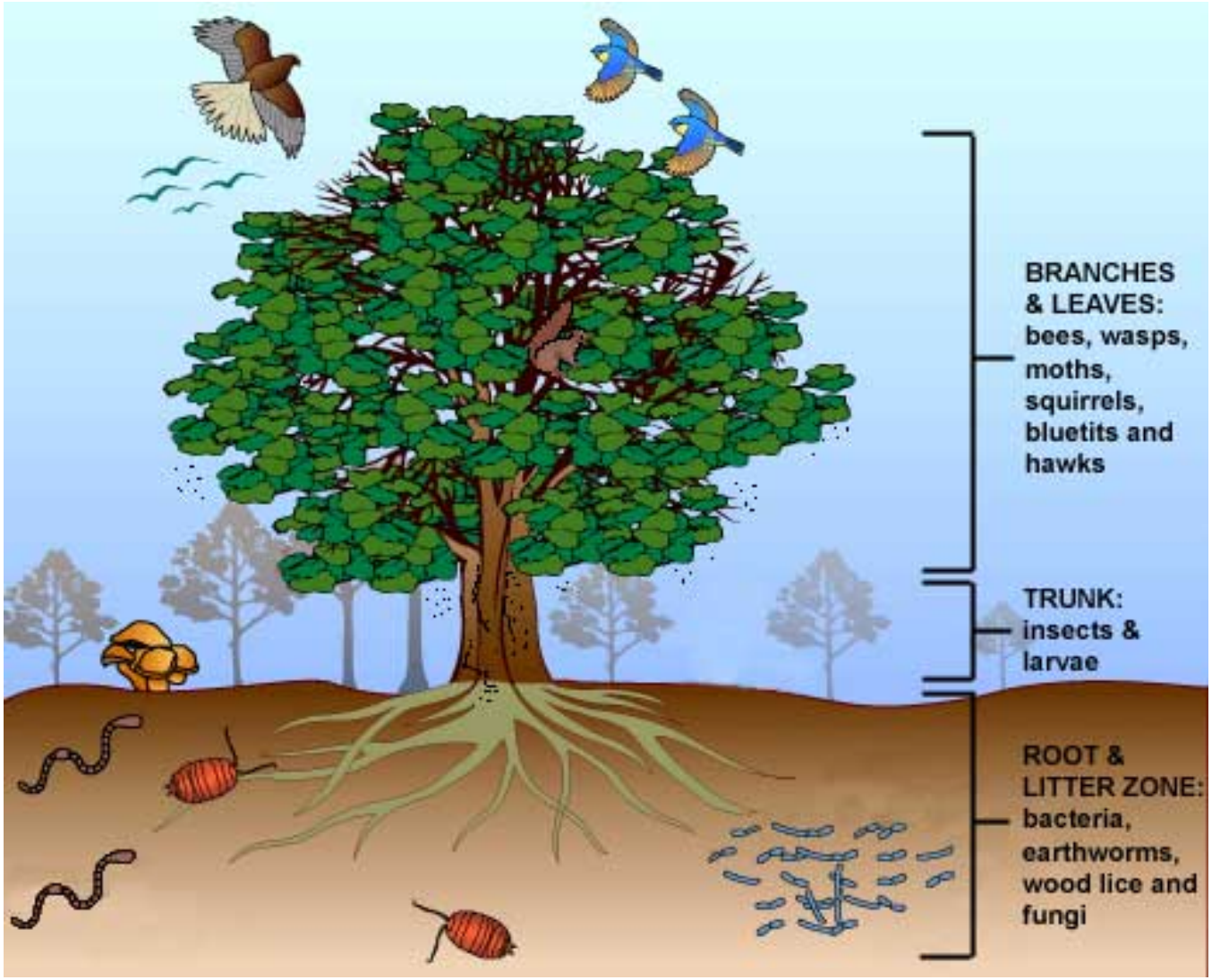
Network sketch with two modules



Origin of transcriptomics complexity?

- Where does the network complexity come from?
  - [In less stable, more changing environments, transcription factors are over-represented] … *This suggests that in ever-changing, highly competitive environments, there is a strong selective pressure towards regulated and coordinated gene expression, compared with very stable environments.* (Cases *et al.*, 2003)
- According to this view, the origin of (transcriptomic) complexity is another complexity (environmental)!
  - But in our experiments, the complex network emerges in a simple environment (one stable state) as well as in two-states environments
- Thus complexity emerges "for free" (at least in the model)
  - Environmental complexity is NOT a necessary condition
  - A new analysis paradigm for genetic networks understanding?
  - What is the environment of an organism? (→ question 3)

Third step: question our perception



http://team.inria.fr/Beagle - http://www.aevol.fr