

Computational Biology II

What is computational biology?

Discipline that uses mathematical/computational techniques to address questions in biology.

Computational biology in practice

Science 9 June 2000:

Vol. 288 no. 5472 pp. 1789–1796

DOI: 10.1126/science.288.5472.1789

[< Prev](#) | [Table of Contents](#) | [Next >](#)

RESEARCH ARTICLE

Timing the Ancestor of the HIV-1 Pandemic Strains

B. Korber^{1,2,*†}, M. Muldoon^{2,3}, J. Theiler¹, F. Gao⁴, R. Gupta¹, A. Lapedes^{1,2}, B. H. Hahn⁴, S. Wolinsky⁵ and T. Bhattacharya^{1,†}

[+ Author Affiliations](#)

ABSTRACT

HIV-1 sequences were analyzed to estimate the timing of the ancestral sequence of the main group of HIV-1, the strains responsible for the AIDS pandemic. Using parallel supercomputers and assuming a constant rate of evolution, we applied maximum-likelihood phylogenetic methods to unprecedented amounts of data for this calculation. We validated our approach by correctly estimating the timing of two historically documented points. Using a comprehensive full-length envelope sequence alignment, we estimated the date of the last common ancestor of the main group of HIV-1 to be 1931 (1915–41). Analysis of a gag gene alignment, subregions of envelope including additional sequences, and a method that relaxed the assumption of a strict molecular clock also supported these results.

Computational biology in practice

HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time

Alan S. Perelson, Avidan U. Neumann, Martin Markowitz,
John M. Leonard, David D. Ho*

A new mathematical model was used to analyze a detailed set of human immunodeficiency virus-type 1 (HIV-1) viral load data collected from five infected individuals after the administration of a potent inhibitor of HIV-1 protease. Productively infected cells were estimated to have, on average, a life-span of 2.2 days (half-life $t_{1/2} = 1.6$ days), and plasma virions were estimated to have a mean life-span of 0.3 days ($t_{1/2} = 0.24$ days). The estimated average total HIV-1 production was 10.3×10^9 virions per day, which is substantially greater than previous minimum estimates. The results also suggest that the minimum duration of the HIV-1 life cycle in vivo is 1.2 days on average, and that the average HIV-1 generation time—defined as the time from release of a virion until it infects another cell and causes the release of a new generation of viral particles—is 2.6 days. These findings on viral dynamics provide not only a kinetic picture of HIV-1 pathogenesis, but also theoretical principles to guide the development of treatment strategies.

Computational Biology II

What is computational biology?

Discipline that uses mathematical/computational techniques to address questions in biology.

Computational biology vs bioinformatics?

Loosely speaking, the emphasis in bioinformatics is on the life cycle of the data (organization, storage and retrieval, visualization), and in computational biology on biological questions, which are addressed by combining (large-scale) data analysis with mathematical modeling and simulation.

Perspective

The Roots of Bioinformatics in Theoretical Biology

Paulien Hogeweg*

Theoretical Biology and Bioinformatics Group, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands



Abstract: From the late 1980s onward, the term “bioinformatics” mostly has been used to refer to computational methods for comparative analysis of genome data. However, the term was originally more widely defined as the study of informatic processes in biotic systems. In this essay, I will trace this early history (from a personal point of view) and I will argue that the original meaning of the term is re-emerging.

Computational Biology II

Why computational biology?

- Biology is changing from a descriptive to a quantitative science.
- We now know most of the components of cells and organisms; recent techniques allow us to get *high-throughput, quantitative measurements* of these components in different conditions.
- How do we make sense of what is going on?
- How do we infer principles of functional organization?

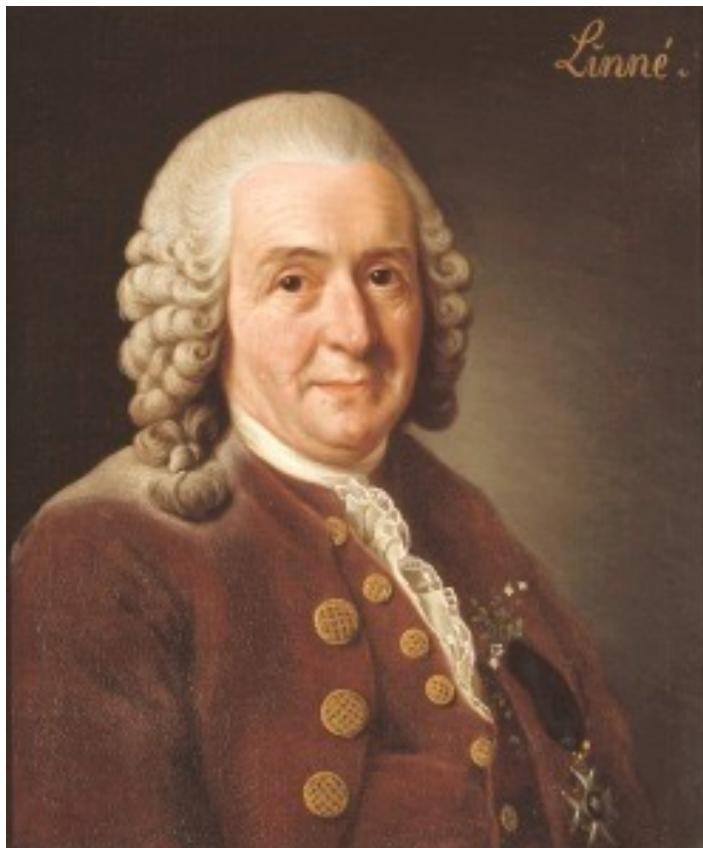
But...

Are there organizational principles in biology?

What do they look like?

Linnaean taxonomy

Carl Linnaeus (1707-1778)
Swedish natural scientist



Systema Naturae, 1735

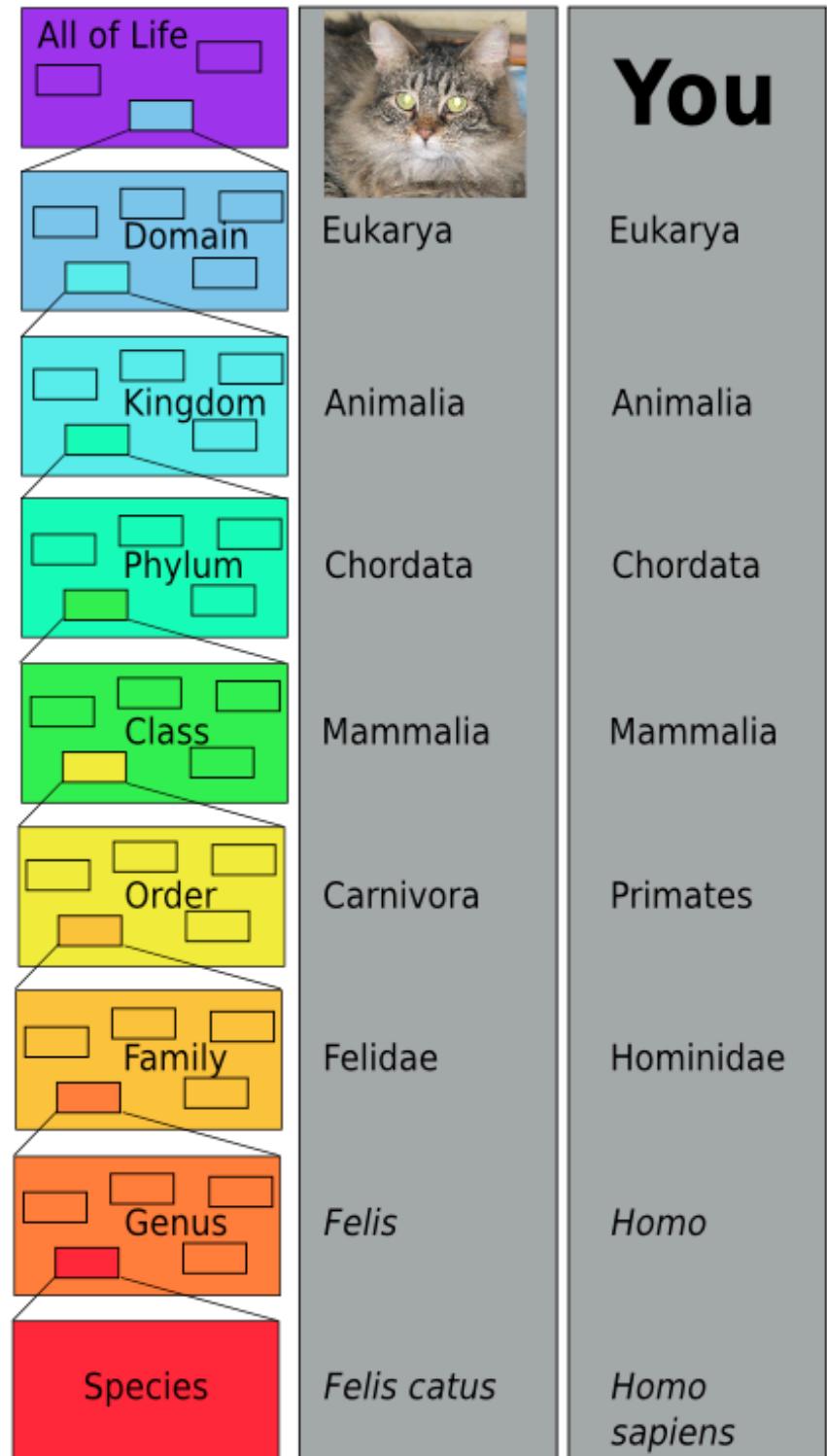
- Organized “nature”
 - living and non-living things
- Standardized the naming of species

Kingdom (Animal, Vegetal, Mineral)
Class
Genus
Species

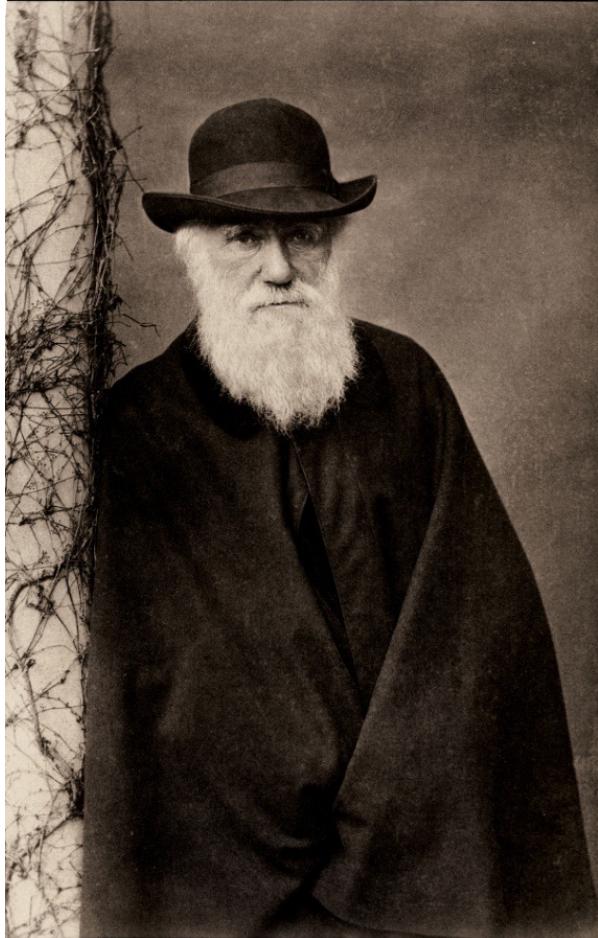
- System describing organisms currently in existence
- Based on morphologic similarity

Taxonomy today

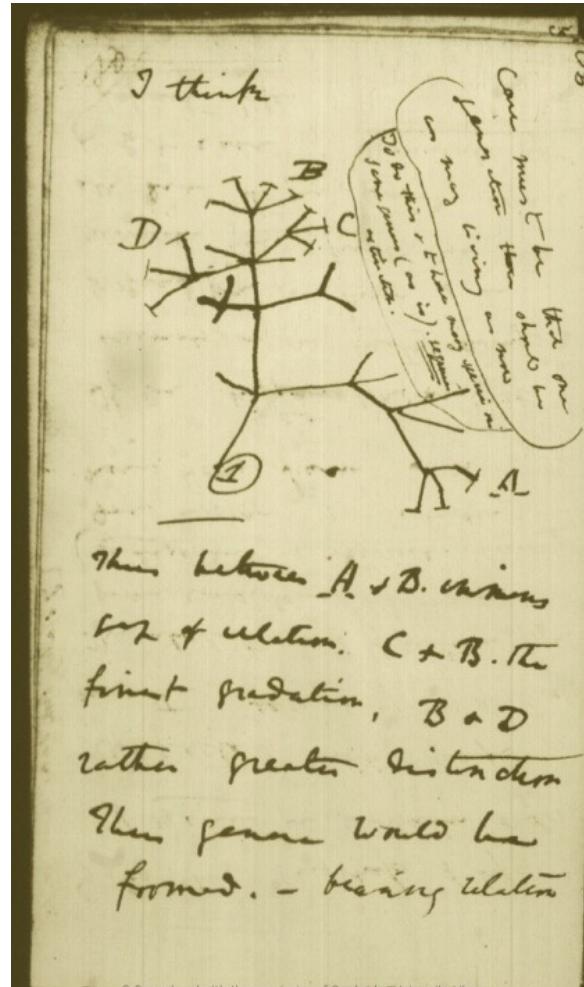
study.com



Darwinian theory of evolution



Charles Darwin (1809-1882)



Darwin's sketch of the tree of life

Fundamental insight:
species that have more
similar traits have a more
recent common ancestor.

Provides a rationale for
the similarity relationship
of species.

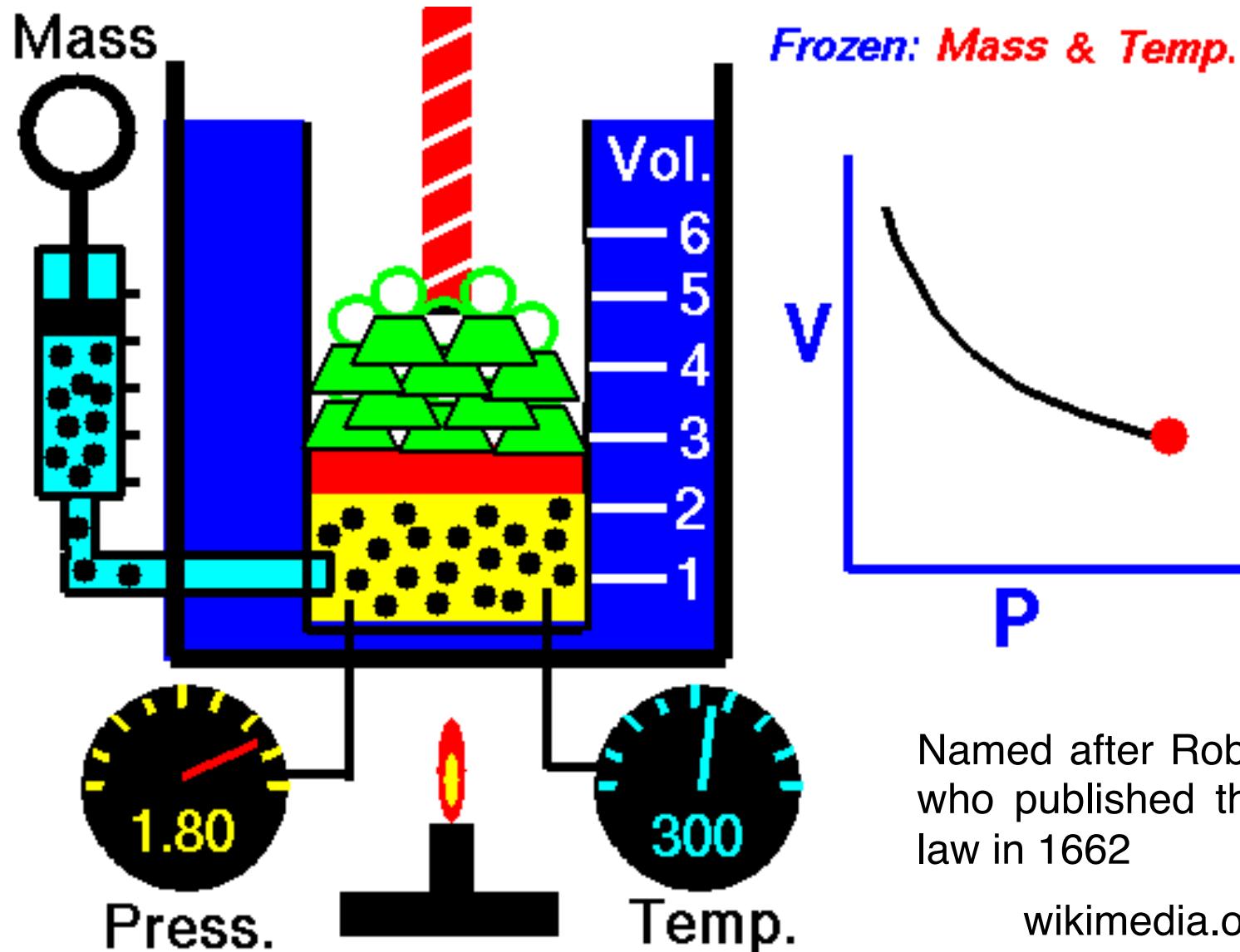
The internal nodes of the
tree represent species that
really existed.

- Big conceptual jump
- Not a quantitative law

Capturing relationships in natural phenomena

Boyle's law

$$P \cdot V = k \text{ if mass and temperature are constant}$$

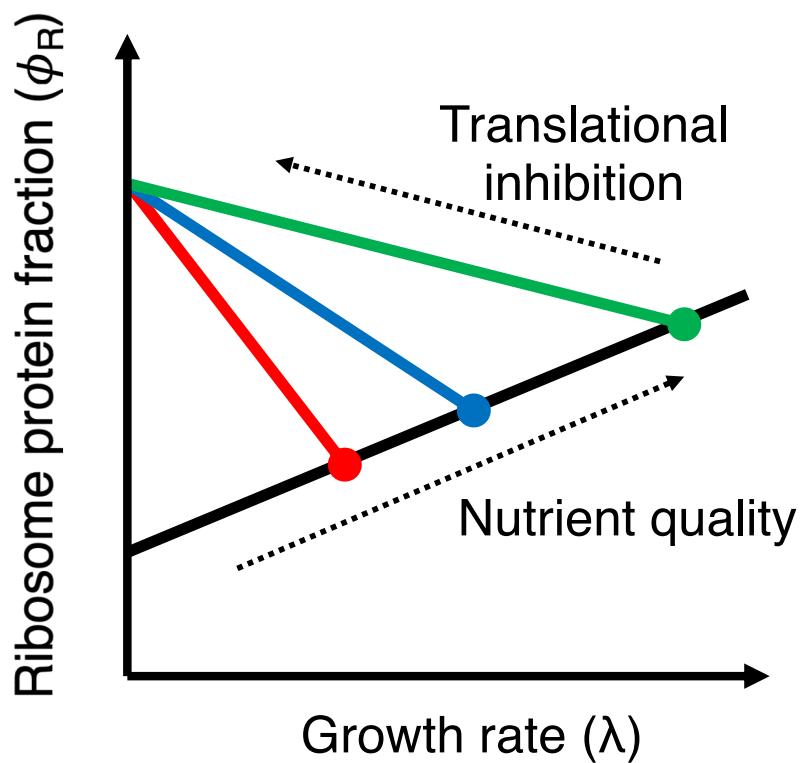


Named after Robert Boyle,
who published the original
law in 1662

wikimedia.org

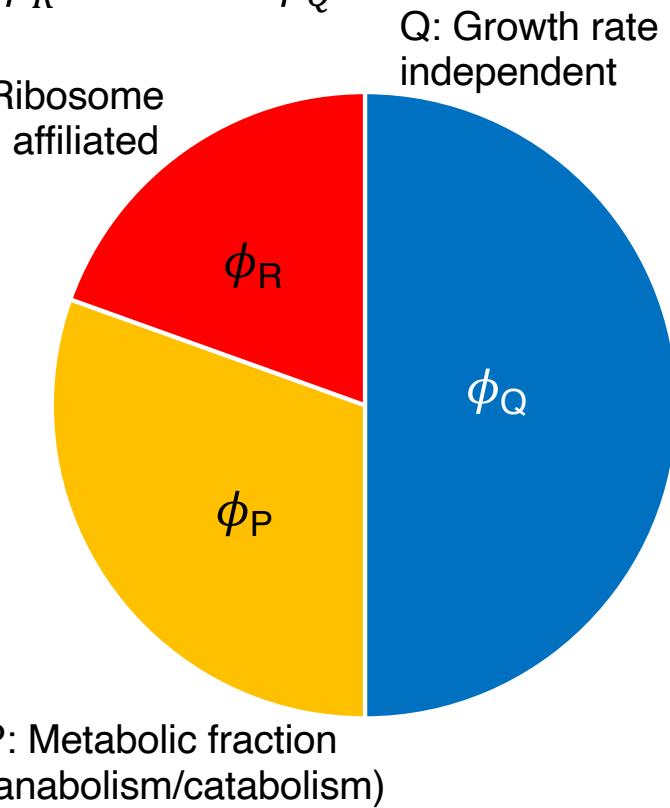
Macroscopic laws in resource allocation

Empirical growth laws in bacteria



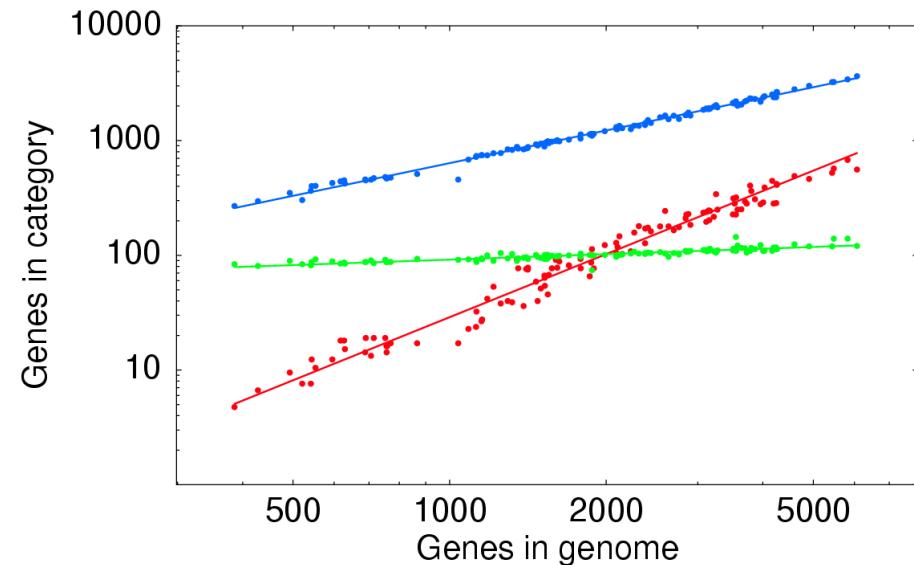
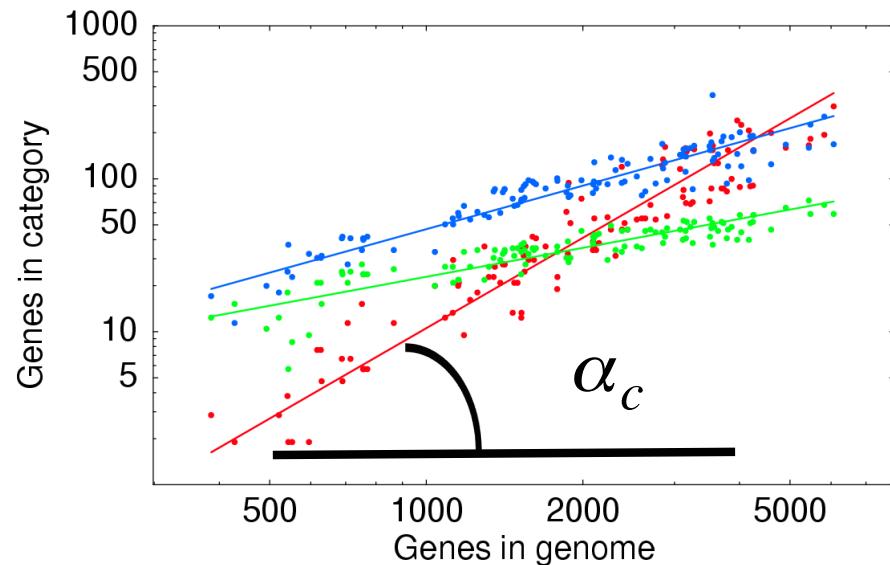
$$\begin{aligned}\varphi_P + \varphi_R &= \varphi_R^{\max} \\ \varphi_R^{\max} &= 1 - \varphi_Q\end{aligned}$$

R: Ribosome
and affiliated



Matthew Scott, Terry Hwa (e.g. MSB 2014)

Empirical laws of genome composition

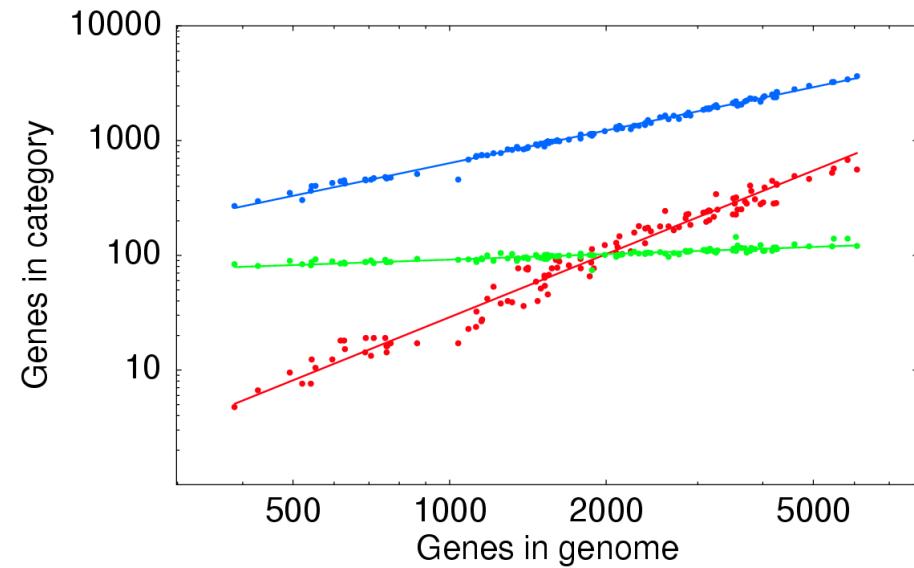
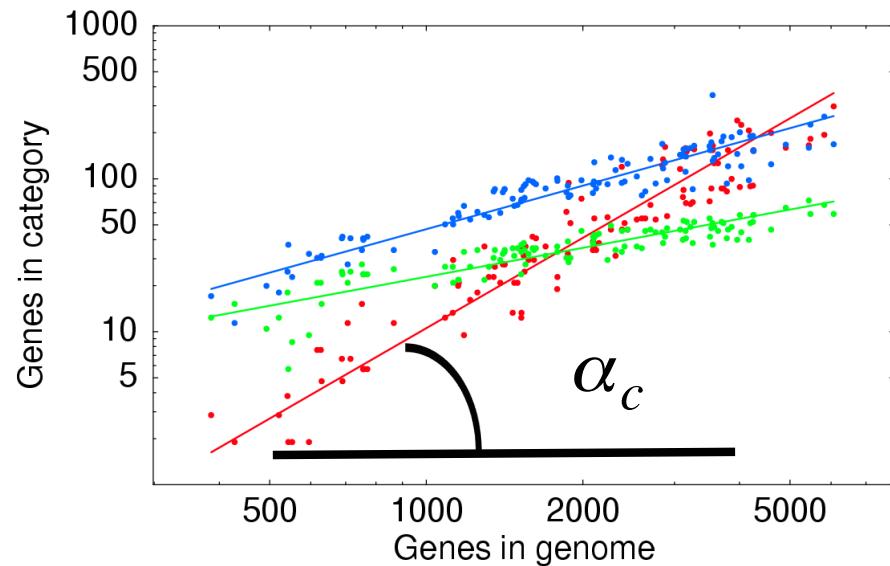


Left panel		Right Panel	
Signal transduction	1.95 +/- 0.15	Transcription regulation	1.83 +/- 0.12
Carbohydrate metabolism	0.94 +/- 0.1	Biological process	0.94 +/- 0.04
DNA repair	0.63 +/- 0.07	Protein biosynthesis	0.15 +/- 0.03

$$n_c = A_c n^{\alpha_c} \Leftrightarrow \log(n_c) = C_c + \alpha_c \log(n)$$

van Nimwegen E. Trends in Genetics (2003)

Empirical laws of genome composition



Left panel		Right Panel	
Signal transduction	1.95 +/- 0.15	Transcription regulation	1.83 +/- 0.12
Carbohydrate metabolism	0.94 +/- 0.1	Biological process	0.94 +/- 0.04
DNA repair	0.63 +/- 0.07	Protein biosynthesis	0.15 +/- 0.03

These relationships tell us something about the organization of living entities and the constraints on it.

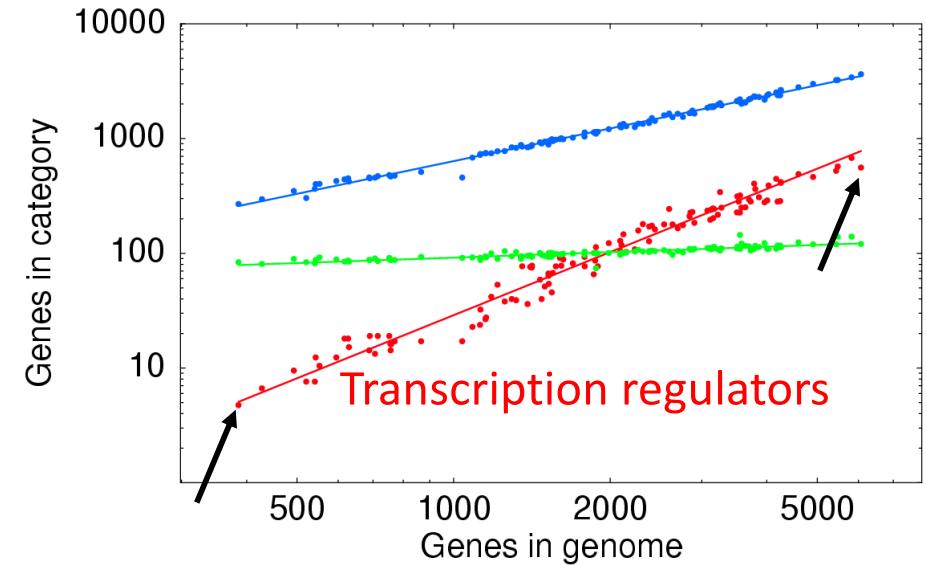
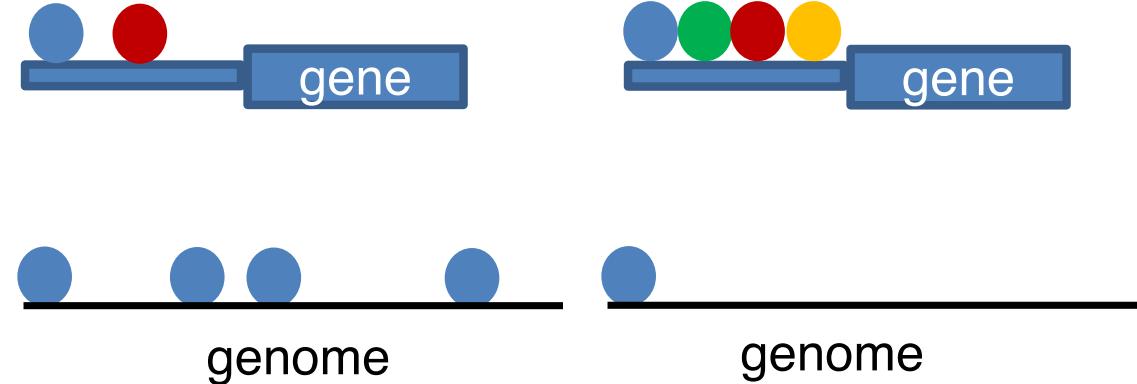
Empirical laws of genome composition

$$n_R \langle \text{targets per regulator} \rangle$$

=

$$n_G \langle \text{regulatory inputs} \rangle$$

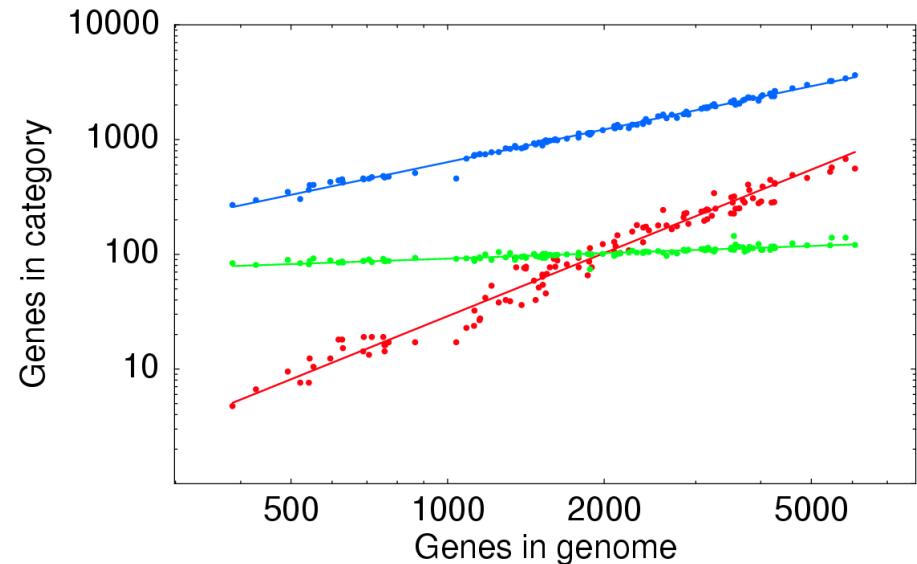
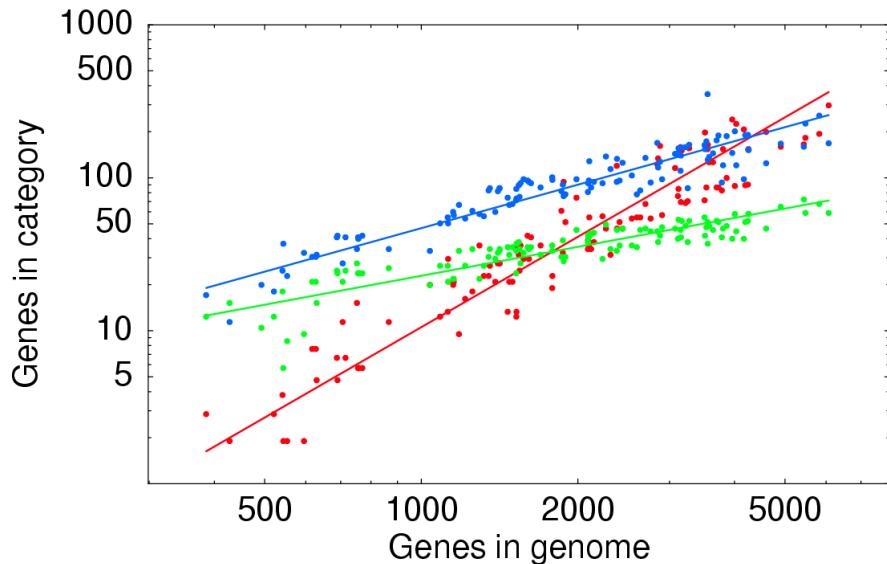
$$\frac{n_R}{n_G} = \frac{\langle \text{regulatory inputs per gene} \rangle}{\langle \text{targets per regulator} \rangle}$$



$$\frac{n_R}{n_G} = \frac{\langle \text{regulatory inputs per gene} \rangle}{\langle \text{targets per regulator} \rangle}$$

$$\frac{n_R}{n_G} = \frac{\langle \text{regulatory inputs per gene} \rangle}{\langle \text{targets per regulator} \rangle}$$

What does it take to generate these plots?



- Genome sequence:
sequencing and assembly of whole genomes
- Gene identification in whole genomic sequences:
gene prediction
- Grouping of genes in functional categories:
functional annotation

Topics that we will address in this course

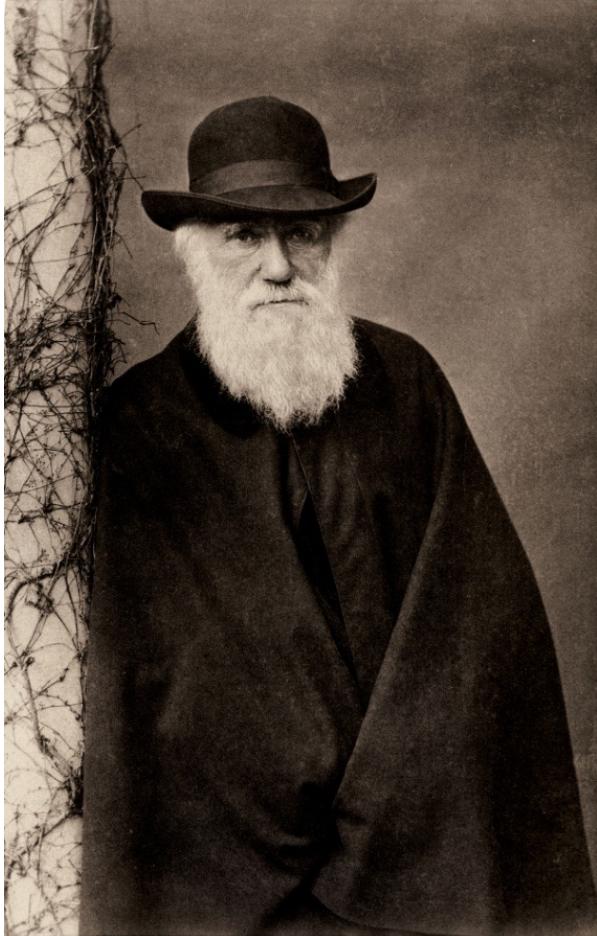
Evolution

An evolutionary process underlies the life we see on Earth today.

What is it evolution?

Evolution is change in the heritable characteristics of biological populations over successive generations.
(Wikipedia)

Evolution



Charles Darwin (1809-1882)

If:

1. entities that reproduce
2. with *heritable* variations
3. affecting reproductive success
4. competing for resources

Then: Evolution!

Any system obeying these 4 conditions will evolve.

Example: MEME = replicating (idea, behavior, style) cultural entity (Dawkins).

- An idea replicates from brain to brain.
- People make small mutations to it as they pass it on.
- Variations on an idea can make it more likely to be passed on. For e.g., if somebody makes a joke more funny by a small alteration, the joke is more likely to be passed on.
- Different memes compete for brain space.

1. Replication

Origin of the building blocks of life (amino acids, nucleotides)

Alexander Oparin
(1924)



- A “primordial soup” of organic compounds can form through the action of sunlight in an atmosphere devoid of oxygen.
- They would form ever more complex molecules until droplets (coacervates) would be generated.
- The droplets could “grow” by fusing with each other and “reproduce” through fission, thereby having a primitive “metabolism”.

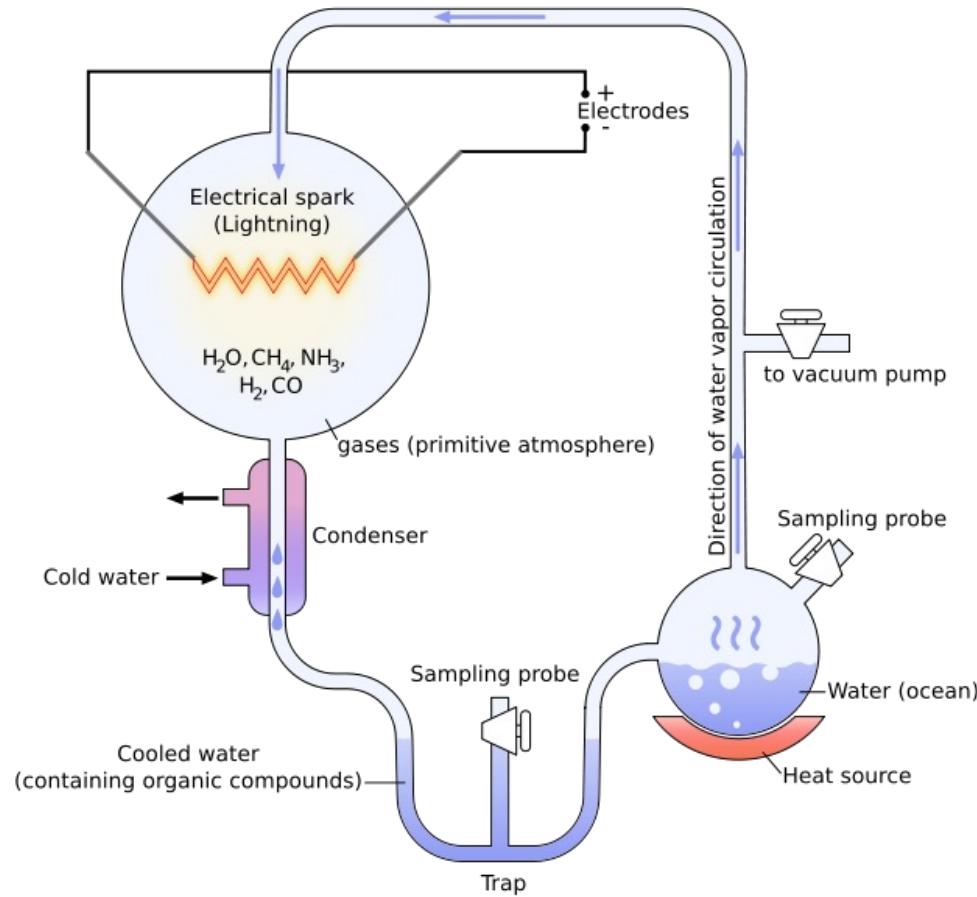
Origin of the building blocks of life (amino acids, nucleotides)

Testing the “primordial soup” hypothesis (1953)



Stanley Miller

Harold Urey



Products: amino acids (mixture of L and D forms, most abundant glycine), sugar, lipids.
Many experiments followed to show that nucleotides can also be synthesized from inorganic compounds under plausible prebiotic conditions.

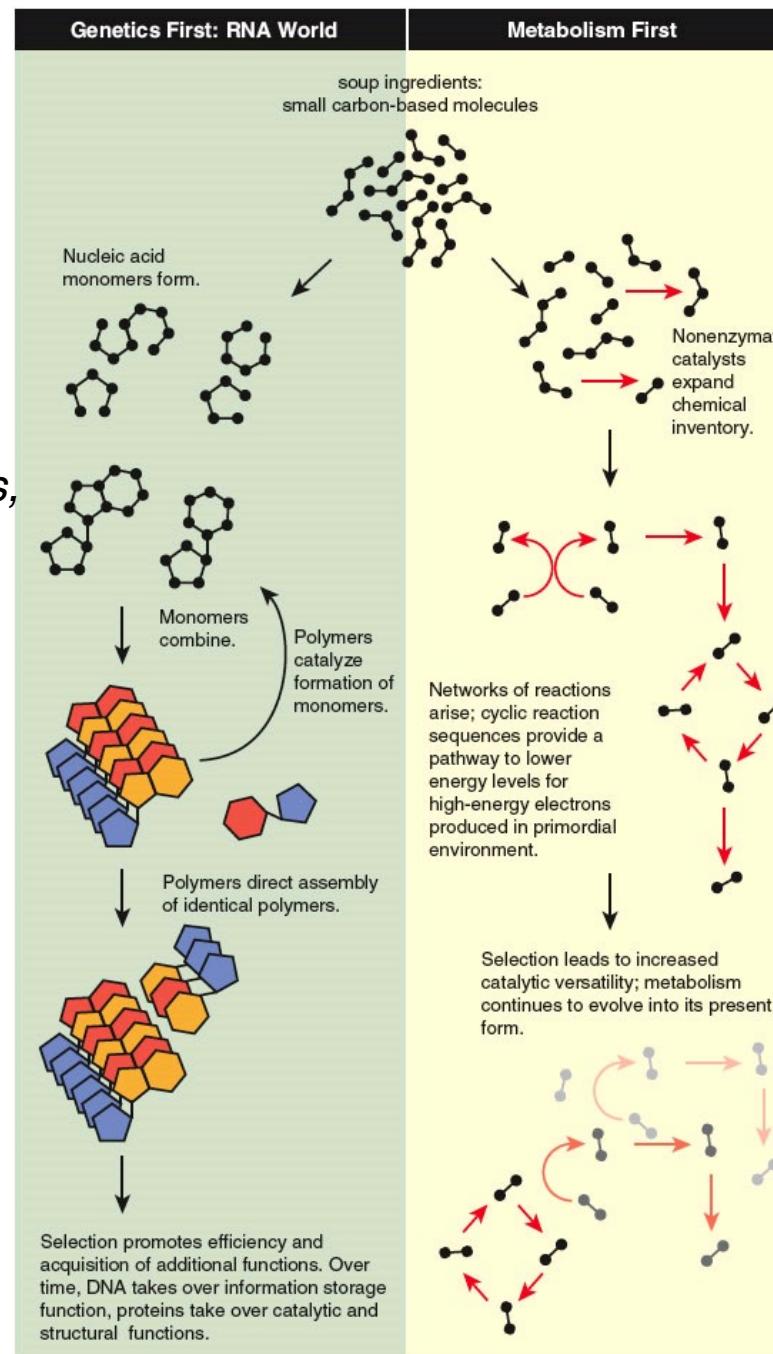
RNA first or metabolism first?

RNA first model.

- From a primordial pool of organic compounds, ribonucleotides are formed.
- Through a chance occurrence, ribonucleotides oligomerize.
- RNAs can act as *enzymes*, including for replicating RNAs.
- Once RNAs are produced that can replicate, evolution starts.

Supporting facts:

- RNAs can both store genetic information and act as enzymes.
- RNAs mediate between DNA and proteins.
- At the catalytic core of the ribosome are rRNAs.



Metabolism first model.

- Many organic compounds available in the primordial soup and many reactions could occur spontaneously (for example at hot vents).
- Eventually an auto-catalytic *cycle* of reactions could be formed.
- Remnants of this cycle would stay at the core of all metabolic networks in organisms (prime candidate: citric acid cycle).
- RNA and genetic information were added later.
- ATP and GTP are universal mediators of energy in all cells.

From:

James Trefil, Harold Morowitz, Eric Smith
American Scientist (2009)

RNA first or metabolism first?

Some issues with the RNA world hypothesis

- cytosine is insufficiently stable at “prebiotic conditions” ($t_{1/2}$ 19 days at 100°C, 17'000 years at 0°C)
- ribose is also unstable under such conditions (73 min at pH 7.0 and 100°C and 44 years at pH 7.0 and 0°C)

Potential solutions

- peptide nucleic acid (PNA), threose nucleic acid (TNA), glycol nucleic acid (GNA)

[Philos Trans R Soc Lond B Biol Sci.](#) 2011 Oct 27; 366(1580): 2902–2909.

PMCID: PMC3158920

doi: [10.1098/rstb.2011.0139](https://doi.org/10.1098/rstb.2011.0139)

The meaning of a minuscule ribozyme

Michael Yarus*

[Author information ▶](#) [Copyright and License information ▶](#)

This article has been [cited by](#) other articles in PMC.

Go to:

ABSTRACT

The smallest ribozyme that carries out a complex group transfer is the sequence GUGGC-3', acting to aminoacylate GCCU-3' (and host a manifold of further reactions) in the presence of substrate PheAMP. Here, I describe the enzymatic rate, the characterization of about 20 aminoacyl-RNA and peptidyl-RNA products and the pathways of these GUGGC/GCCU reactions. Finally, the topic is evolution, and the potential implications of these data for the advent of translation itself.

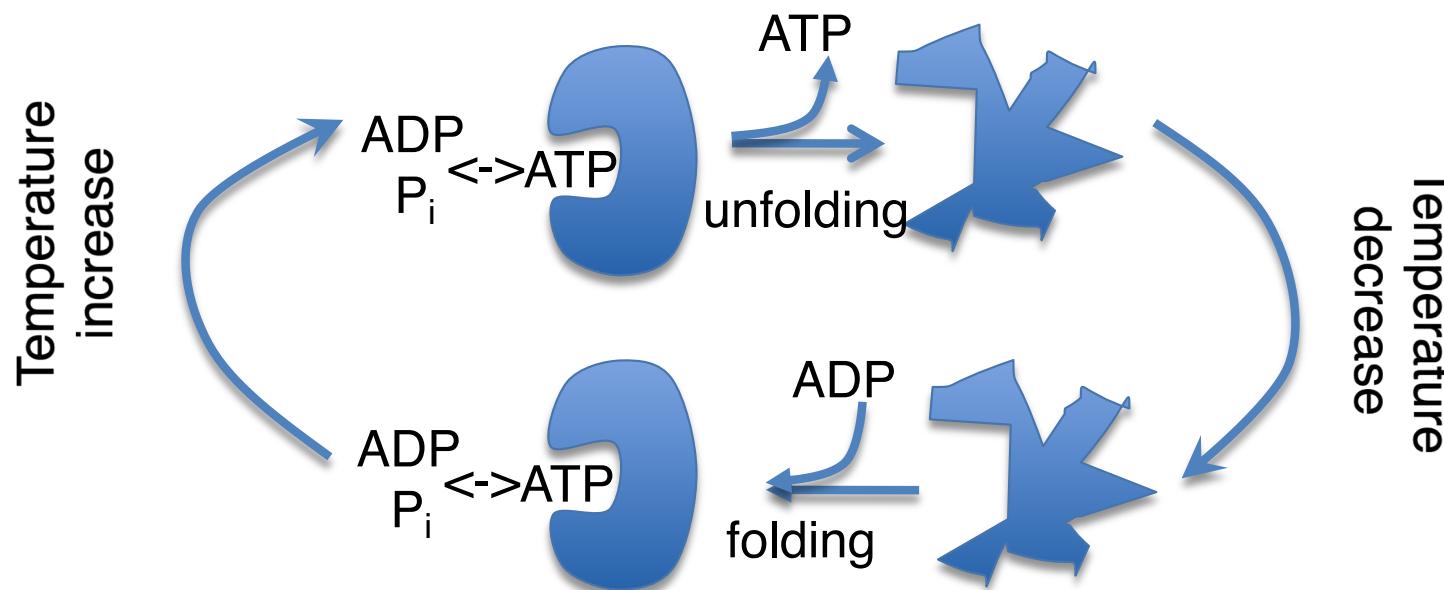
Wächtershäuser's 'Pioneer organism'

Hypothesis: earliest form of life originated in a volcanic hydrothermal flow at high pressure and high (100 °C) temperature.

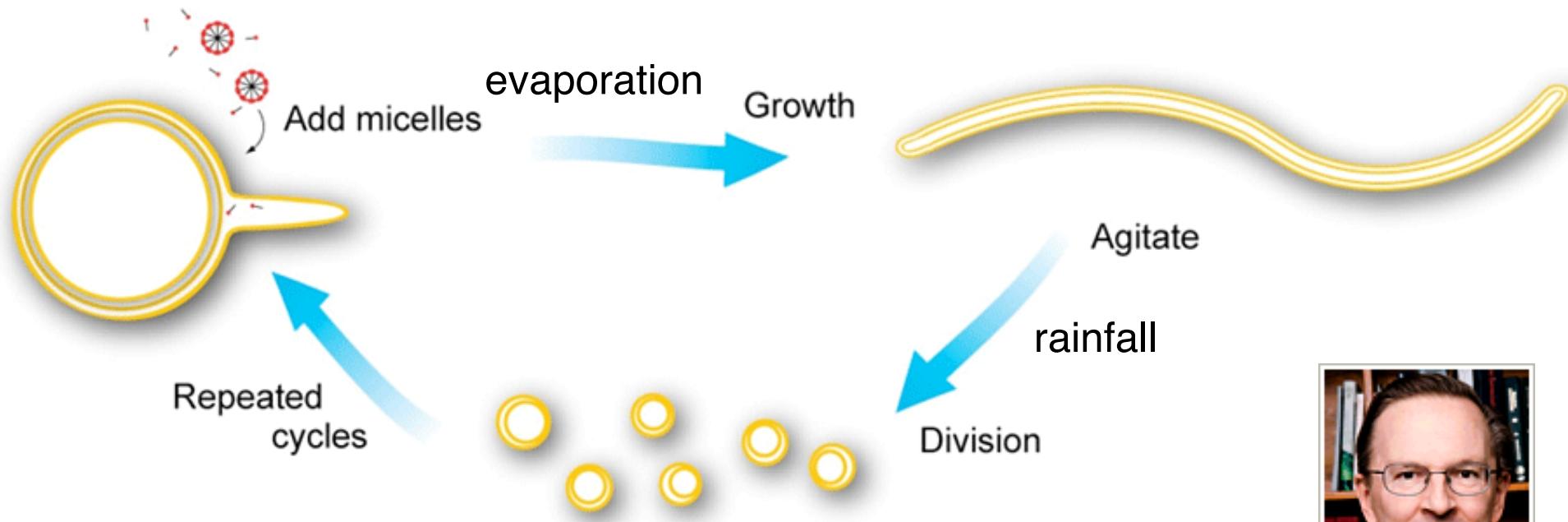
Evolution of the first metabolic cycles (<https://www.pnas.org/content/87/1/200>)

Iron-sulfur world: self-sustained reactions based on the chemistry of thioesters that took place on surfaces (possible connection to the acetyl-CoA synthase of today).

Thermosynthesis world (Anthonie Muller, <https://pubmed.ncbi.nlm.nih.gov/16024164/>): thermal cycling, which could have occurred as the protocell was suspended for e.g. in a hot spring, would have provided the energy for condensation of substrates bound to the “First protein” (possible connection to ATP synthase of today).



Origin of replication in nature: Splitting vesicles



- Starting from a lipid vesicle, addition of micelles leads to spontaneous growth of the vesicle.
- Currents in the solution can easily lead the grown filament to break into pieces.
- This leads effectively to replication of the vesicle.



Jack Szostak
2009 Nobel Prize
Physiology & Medicine

Replication

- One of the key features thought to distinguish ‘living’ things from ‘dead’ things.
- Not so long ago, it was unclear to many whether it would be possible to devise artificial objects that *replicate themselves*.
- The mathematician John von Neumann decided to create an abstract mathematical ‘world’ with simple rules to show how replication can be implemented.

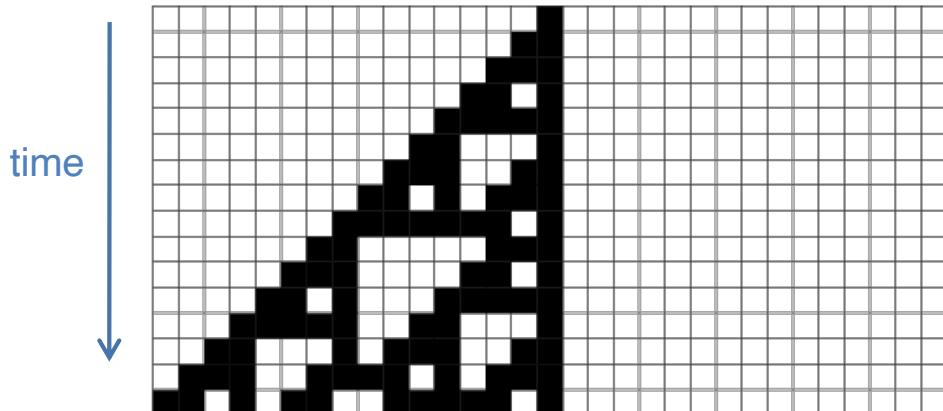
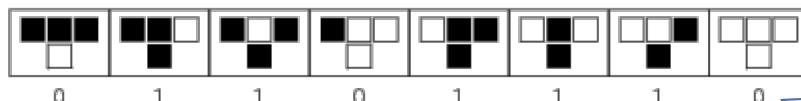
Cellular Automata



John von Neumann
1903-1957

1D 2-state CA

rule 110



Cell plus its neighbors can be in 8 different states.

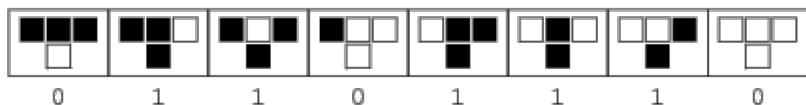
The rule specifies the middle cell’s state at the next time point.

Start from a single cell ‘on’.

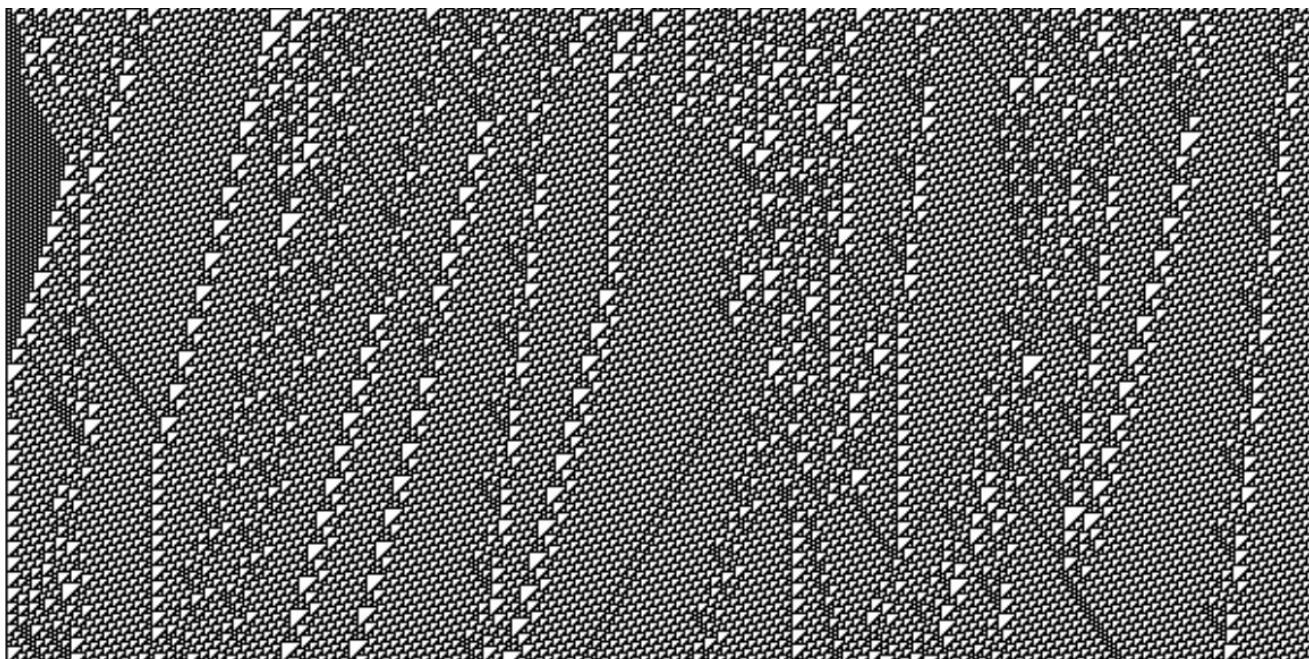
A pattern will develop over time in this ‘world’.

Elementary CA rule 110

rule 110



Behavior of rule 110 starting from random initial conditions.



Shows that very complex dynamics can happen even in a world with such simple rules.

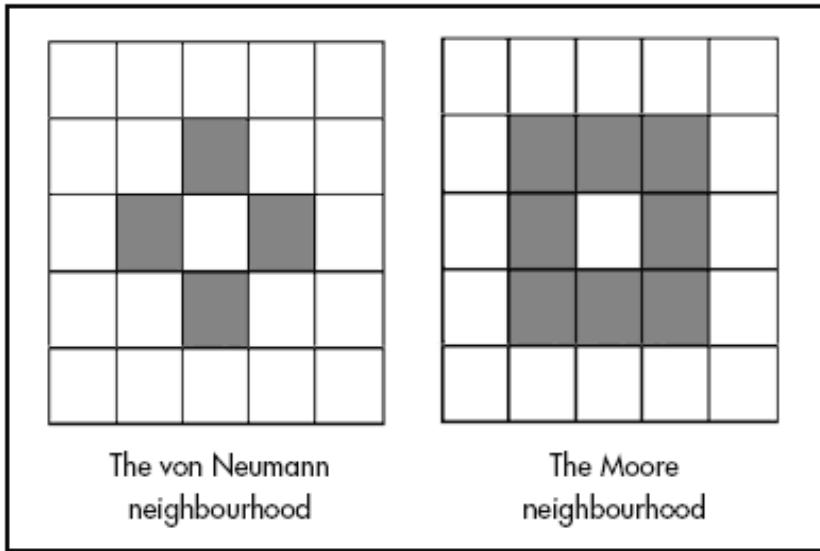
Side note:

Around the year 2000, Matthew Cook, working as an assistant to Stephen Wolfram, proved that rule 110 is ‘Turing complete’, i.e. it can implement *any computation* that any computer can do.

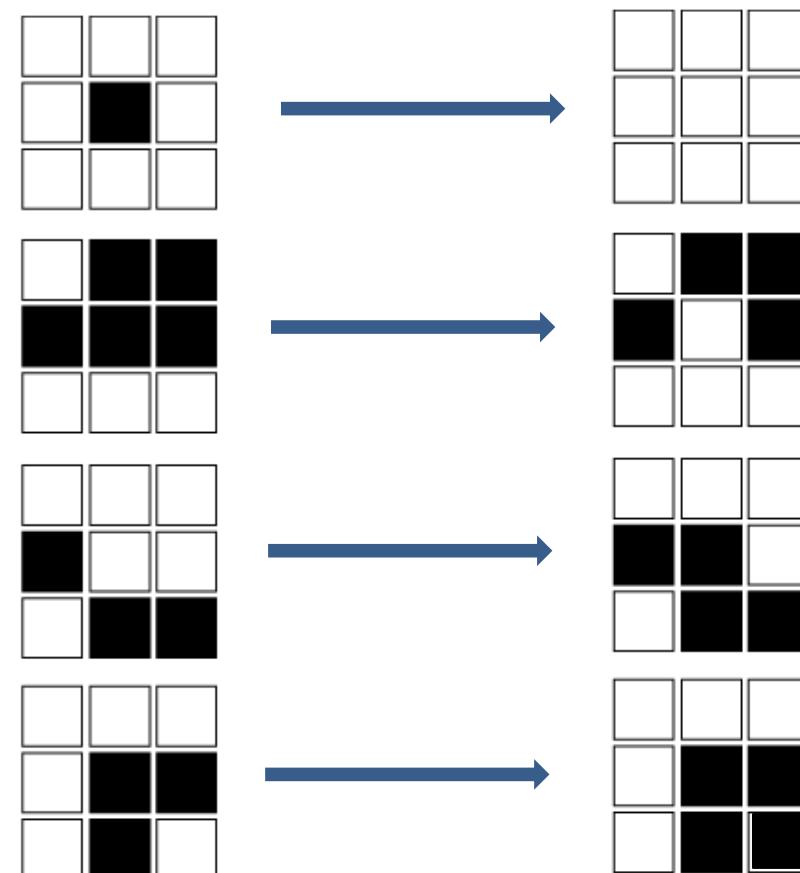
Universality in Elementary Cellular Automata *Complex Systems* 15: 1-40 (2004).

See also: S. Wolfram: A New Kind of Science (2002)

2D Cellular Automata



- The state of each cell at the next time-point is again a function of its state and that of its neighbors.
 - Von Neumann used a neighborhood with 4 orthogonal neighbors.
 - Many CAs use instead the Moore neighborhood with 9 neighbors.



Example:

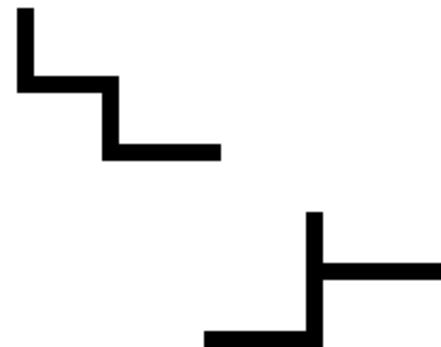
John Conway's Game of Life (1970)

- A cell that is 'on' will stay on if 2 or 3 of its neighbors are on.
- A cell that is 'off' will come 'alive' if precisely 3 of its neighbors are on.

Conway's Game of Life

Simple example:

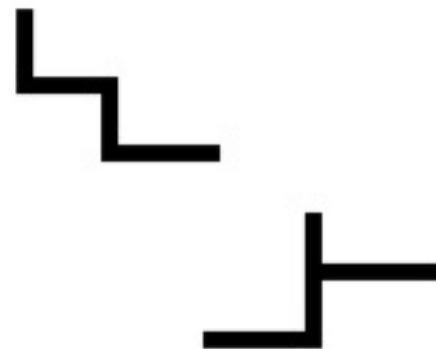
Generation: 1 Creatures: 41



Conway's Game of Life

Simple example:

Generation: 1 Creatures: 41



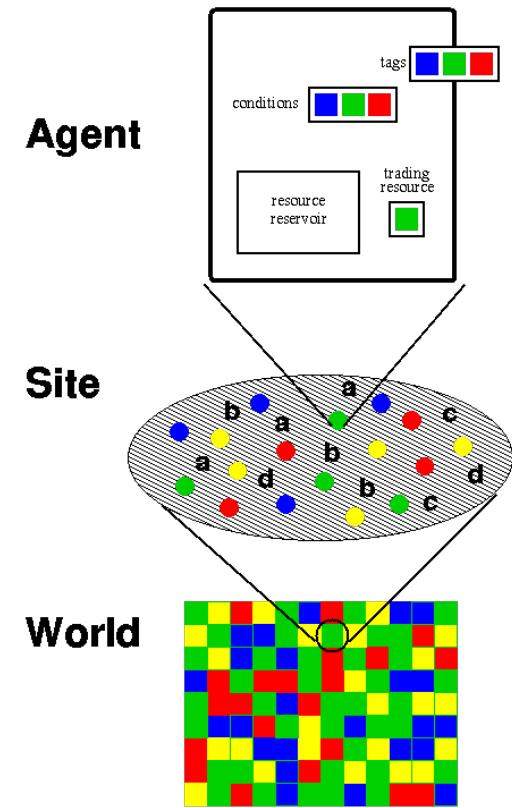
A 'glider gun'. A pattern that spits out moving objects:



Simulating evolution in a computer with CAs

Echo (Holland, Forrest, Jones, Hraber, et al.)

- simulation tool developed to investigate mechanisms that regulate diversity and information-processing in systems comprised of many interacting adaptive agents.
- Echo agents interact via combat, trade and mating and individual genotypes encode rules for interactions.
- populations of these genomes evolve interaction networks that regulate the flow of resources.
- Reemerging in the study of intra-patient cancer evolution
- Bioinspired computing: genetic algorithms as a problem solving approach



2. Heritable variation

Fuel of evolution

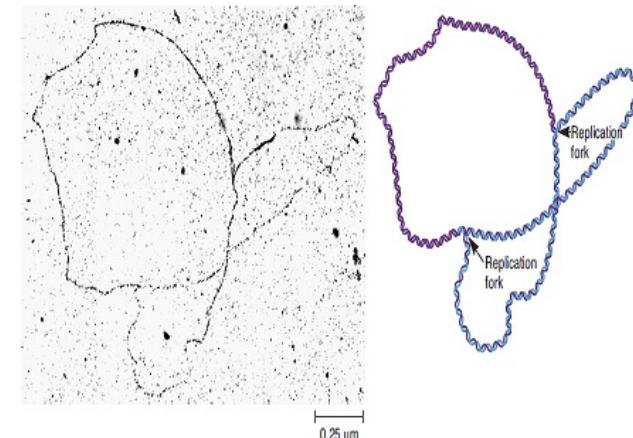
Genetic variation

chegg.com

Where do mutations come from?

The DNA can get damaged.

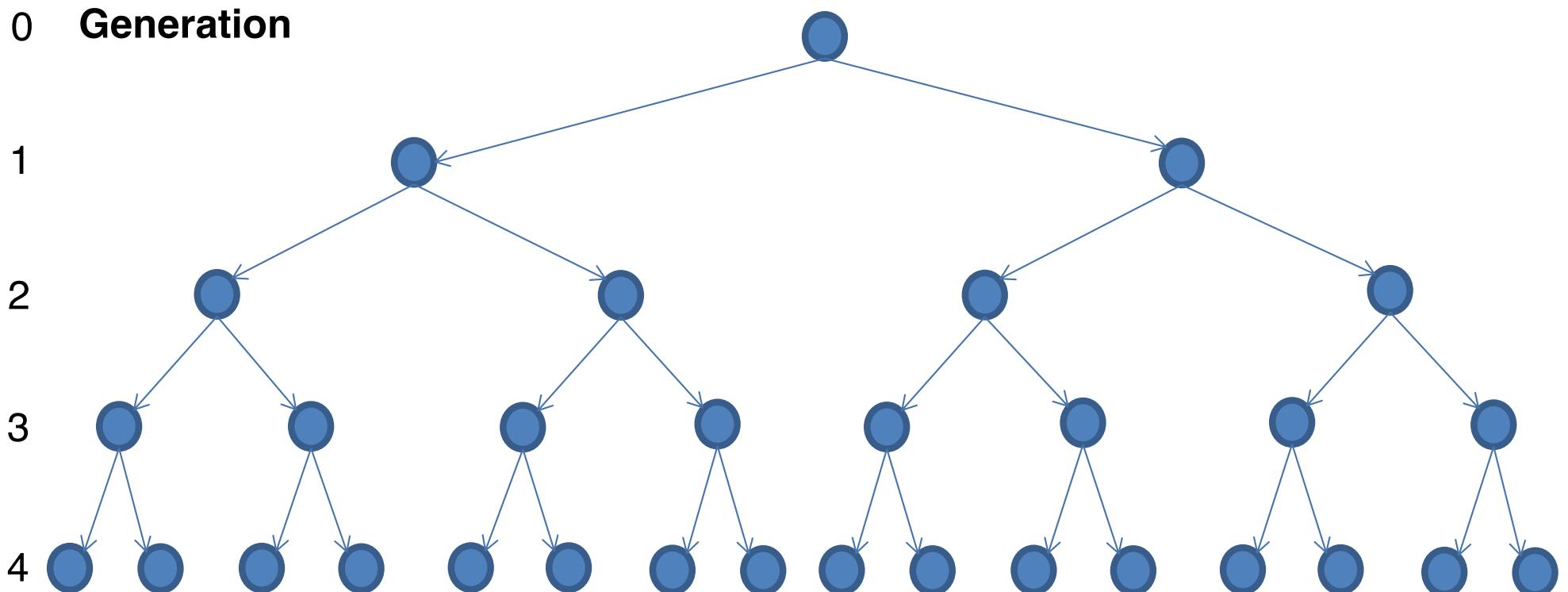
DNA replication is error-prone.



- Replication and repair are done by tiny molecular machines
- Which are subjected to thermal ‘noise’, other molecules bumping in to them, polymers flopping around, twisting, turning, shaking, etc.
- Mistakes happen
- It is almost impossible to predict where what error will occur. Therefore we describe these *mutation* events as *random*
- This does not mean that mutations don’t have some kind of bias (to keep in mind: mutational hot spots)

Simplest family relationships: Trees

Let's look at an idealized dynamics of a bacterial cell culture.



- Each generation the number of individuals doubles
- If we draw arrows from the 'parent' individual to the 'offspring' we obtain a *tree* structure

Heritable variation

0 **Generation**

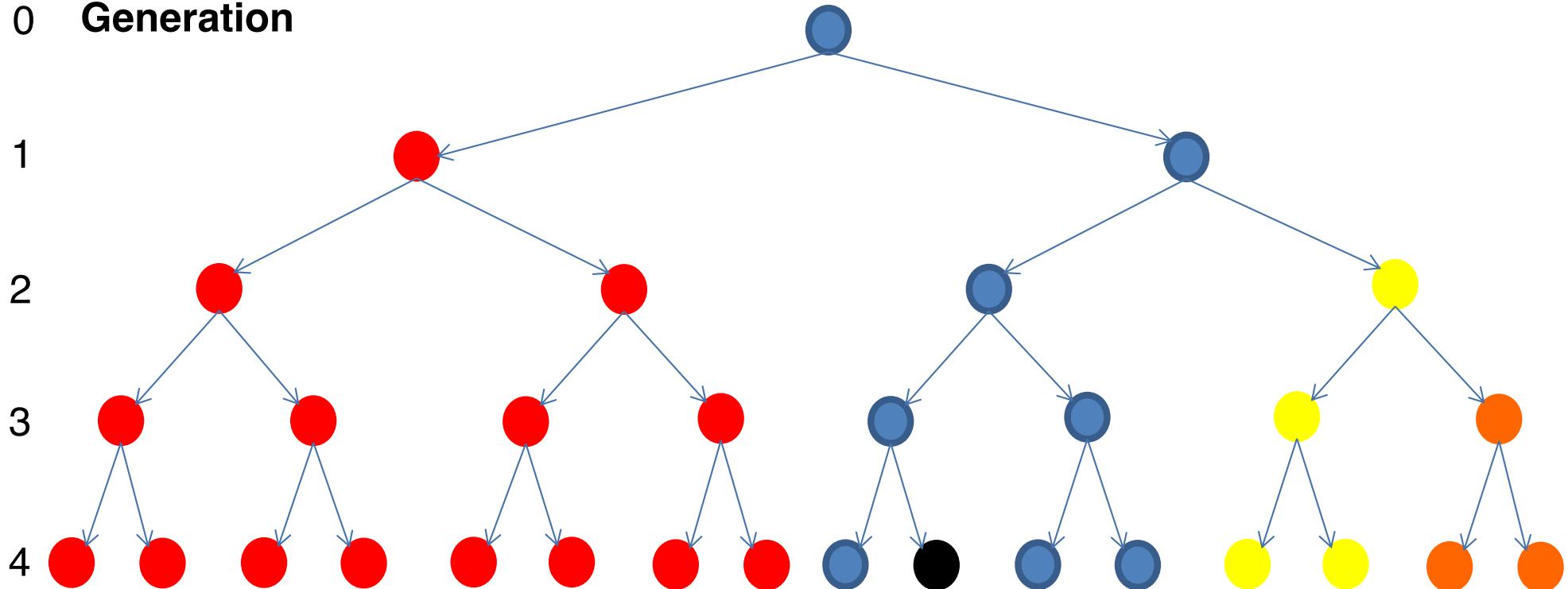
1

2

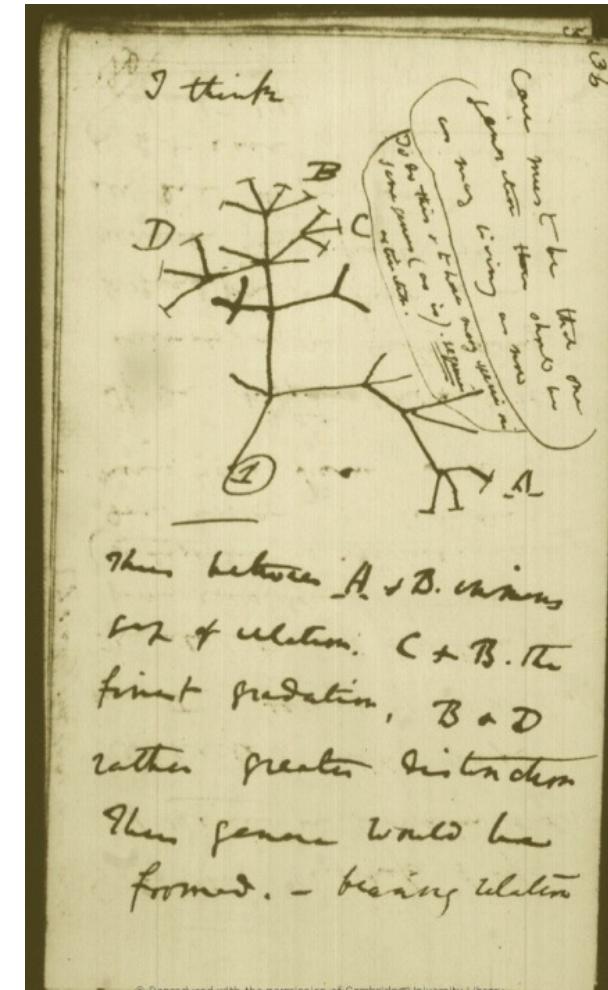
3

4

- Each generation the number of individuals doubles
- If we draw arrows from the ‘parent’ individual to the ‘offspring’ we obtain a *tree* structure
- Heritable ‘mutations’ naturally induce *hierarchical* similarities, reflecting *family relationships*



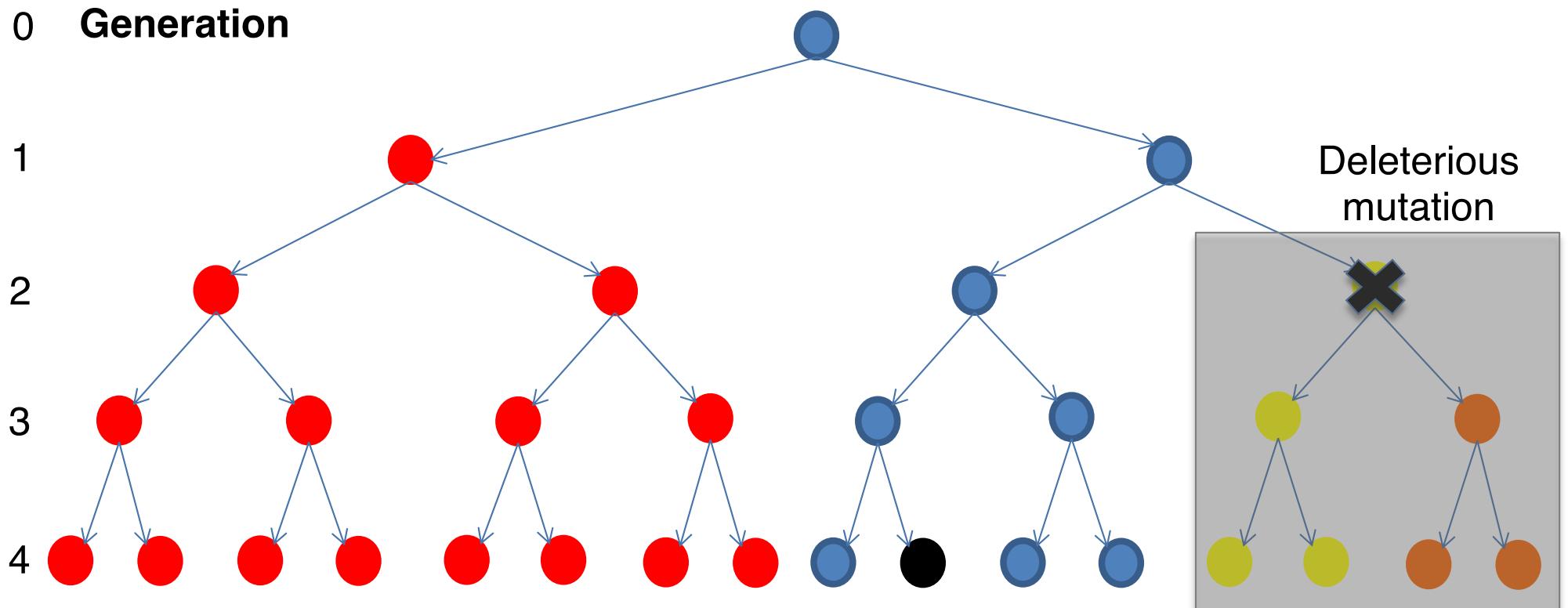
Consistent with species classifications



But very different meaning

3. Reproductive success

Reproductive success



In our initial system
each bacterium divided into two
the time between divisions was identical

More realistically, mutations affect
whether and how fast bacteria are dividing

The structure of the resulting
trees will be very different

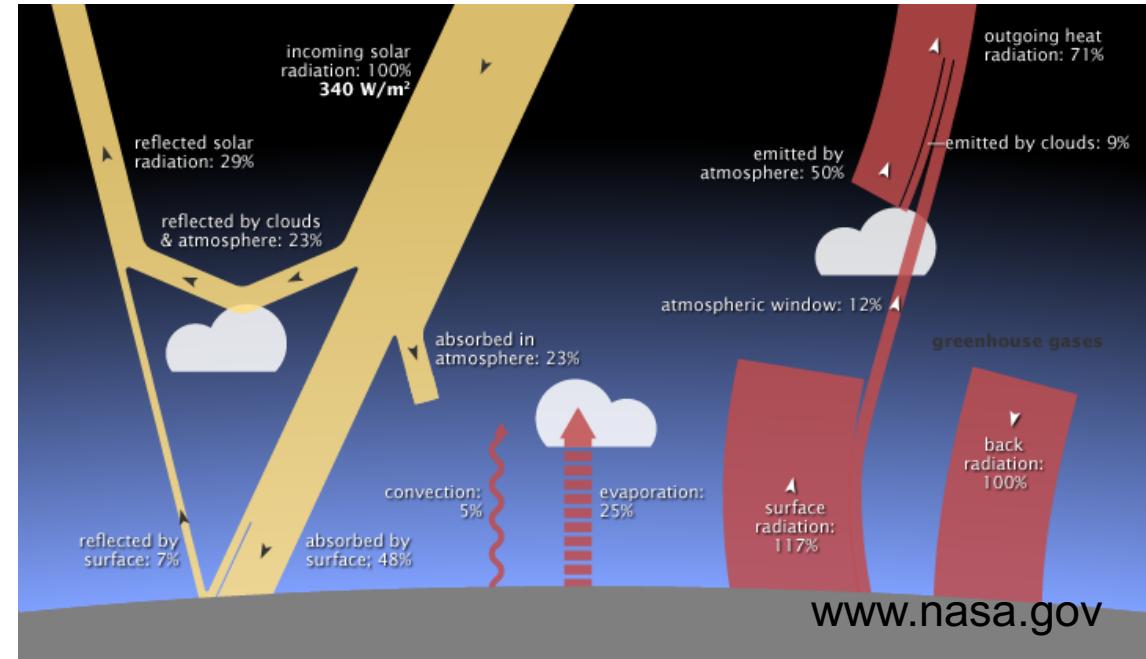
4. Competition for resources

Competition for resources. Energy flows



Physics tells us
*energy is a
conserved quantity.*

To replicate, a
structure needs
energy from
somewhere.



- Essentially all energy input to the planets comes from the sun.
- Where does this energy go?

The energy will *dissipate* through the system.

- The molecules in the atmosphere and on the surface of the planet will *heat up*.
- Differences in heat can cause *currents* of gases and liquids.
- These currents cause erosion of materials, and drag small particles along.
- Provided enough heat is present *chemical reactions* will take place.
- This can *alter* the state of the planetary system.
- Many of these events occur in cycles, hot/cold, light/dark, etc.
- Thermodynamics: Energy will go wherever it can.

Energy flows. Competition for resources

One way to think of evolution:

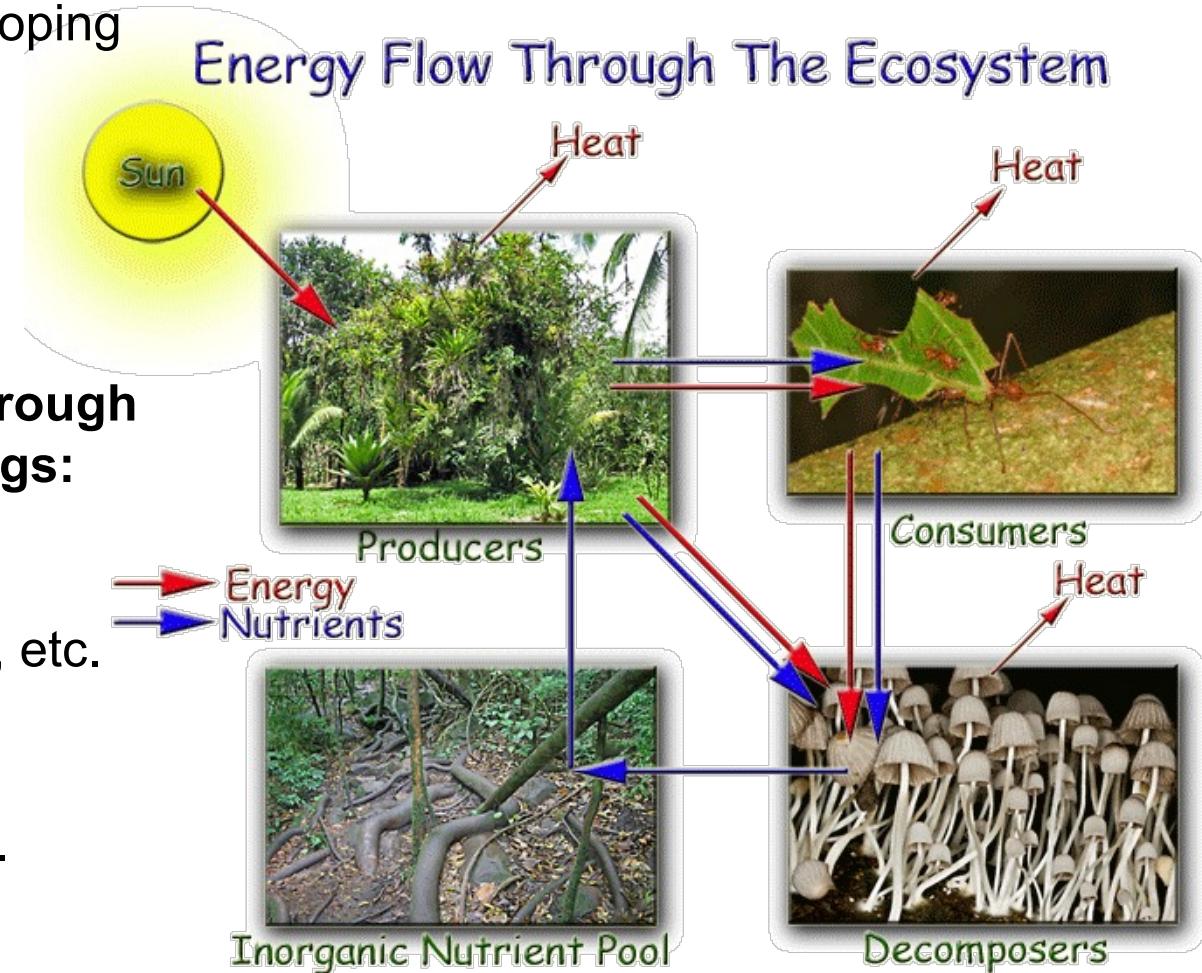
- Evolution is a very slowly developing response of the physical system 'earth' to energy from the sun. It constantly finds more elaborate 'paths' for dissipating energy.

Energy dissipates further through the ecosystem of living beings:

- Plants use sunlight directly.
- Ants eat leafs.
- Fungi decompose dead ants, etc.

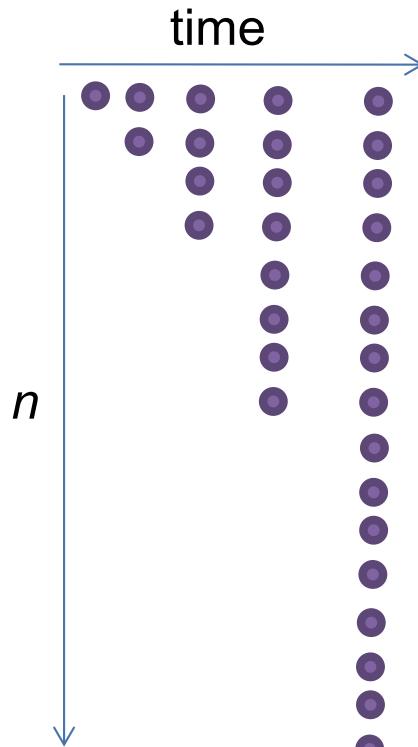
Other conserved quantities:

- Water (more or less recycled).
- The total number of atoms of different kinds, i.e. important atoms like carbon, oxygen, nitrogen, and phosphate cannot be replicated, they have to be taken from somewhere.



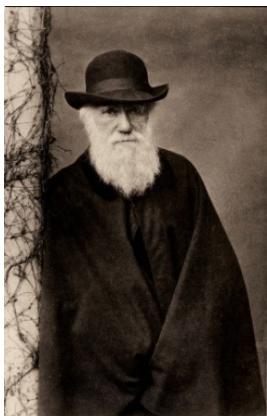
Competition

The number of replicating entities grows *exponentially* in time



$$n(t) = 2^{t/\tau} \quad \text{with } \tau \text{ the replication time.}$$

- The amount of energy and basic components needed are thus also growing exponentially.
- Very quickly, there will not be enough resources (basic compounds and energy) to keep doubling.
- To keep growing in number, there are two options for a replicator:
 1. Find a new source of energy and compounds, i.e. find a new *niche* to live in.
 2. Get *better* at getting the resources than its neighbor, i.e. *outcompete* them.



Crucial ideas:

Limited resources imply that there will be a competition between replicators and this will lead to evolution of 'fitter' individuals.

When replicators find ways to use different resources this will lead to co-existence of replicators in different *niches*.