Is biology more than stamp collecting?

Torgeir R. Hvidsten
Assistant professor in bioinformatics
Umeå Plant Science Centre (UPSC)
Computational Life Science Cluster (CLiC)

Ernest Rutherford

"All science is either physics or stamp collecting"

Outline

- Rutherford’s statement in a scientific and historic context
- Complexity and reductionism in science
- What do I mean by “*Is biology more than stamp collecting?*”?
- Evolution, the computer and design principles in biology
- Laws of genome evolution
- Conclusion: Is biology more than stamp collecting?

What is science?

- The process of evaluating proposed models of nature against observed data
  - Physics, chemistry and biology, but not mathematics or computer science

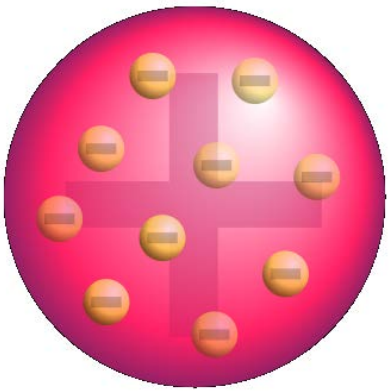
- Mathematics: a theory gains immortality by mathematical proof
  - Fermat's last theorem: $a^n + b^n = c^n$ have no integer solution for $n > 2$
- Science: an experiment consistent with the model only add further evidence to it's validity
  - *all swans are white*
    - Hypothesis (unsupported)
    - Theory (empirical support)
    - Law (massive empirical support)
    - Rejected (by a single experiment)

Historic context

Ernest Rutherford (Lord Rutherford of Nelson)
30 August 1871 – 19 October 1937

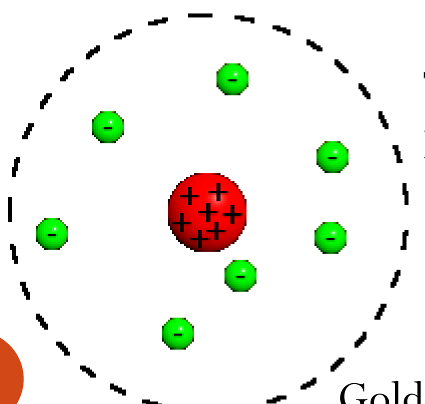
New Zealand-British physicist

"The father of nuclear physics"



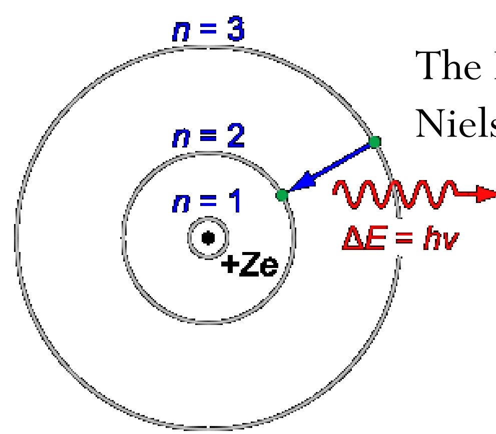
The plum pudding model of the atom
J. J. Thomson in 1904

"If your experiment needs statistics, you ought to have done a better experiment".
Ernest Rutherford



The atomic planetary model
E. Rutherford in 1911

Gold foil experiment (1907)

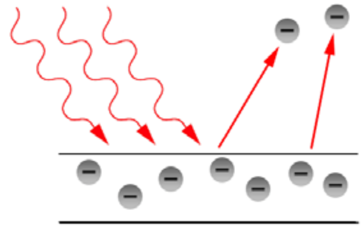


The Bohr model
Niels Bohr in 1913

"Complexity" of the sciences

Physics

Photoelectric effect

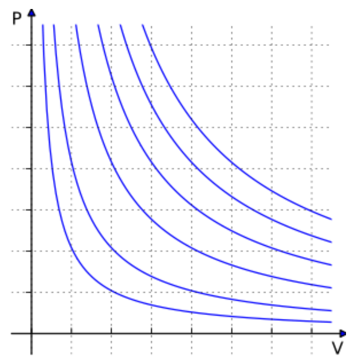


$$E = h\nu$$

E = energy
H = Planck's constant
ν = frequency of light radiation

Chemistry

Ideal gas law

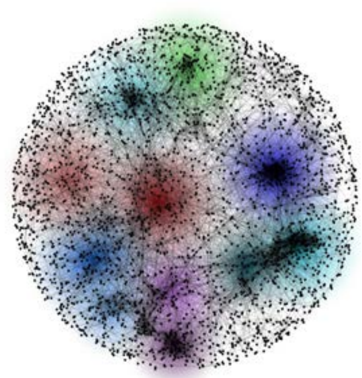


$$PV = nRT$$

P = absolute pressure
V = volume of the vessel
n = number of moles of gas
R = ideal gas constant
T = absolute temperature

Biology

Gene interactions



$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$

y_i = gene expression of gene i
n = number of genes
α = transcription rate
β_{ij} = effect of gene j on gene i

Overfitting and Occam's razor

Constraints from experiments (data):

Model parameters (fitted):

$$x = 7y$$
$$y = 3 + x$$

Has a unique solution: x=-3.5, y=-0.5

$$x = 7y$$
$$y = z + x$$

Has many solutions:

z=3, x=-3.5, y=-0.5
z=6, x=-7, y=-1
…

$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$

n=~30 000 for humans

n > 20 000 for plants

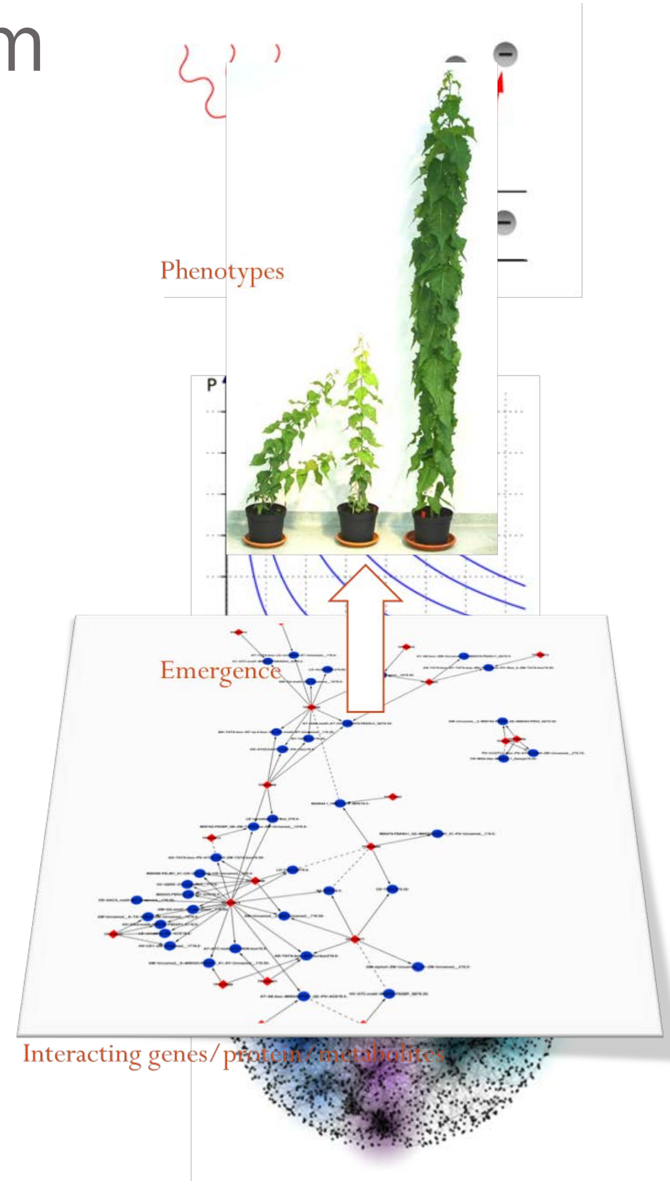
Occam's razor: The simplest model that best explains the data should be chosen

- Require weighting model complexity (no. parameters) against model fit (p-value)
- Example: multiple hypotheses correction (significance threshold = 0.05/n)

Holism versus reductionism

- Can understanding of biological systems be reduced to understanding of individual genes, proteins and metabolites?
  - Systems biology: phenotypes emerge from interacts between genes, proteins and metabolites

- The bigger picture (Vienna Circle, 1922):
  - Inter-discipline reductionism
  - Can biology be reduced to chemistry?
  - Can chemistry be reduced to physics?



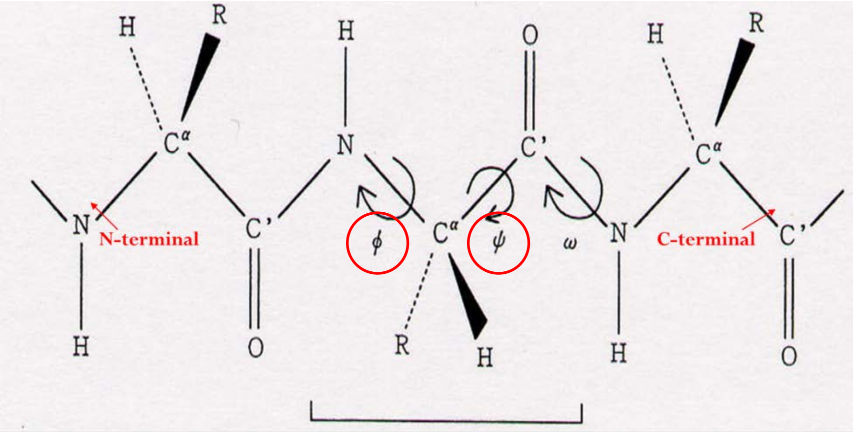
Gatherer, BMC Systems Biology 4:22, 2010

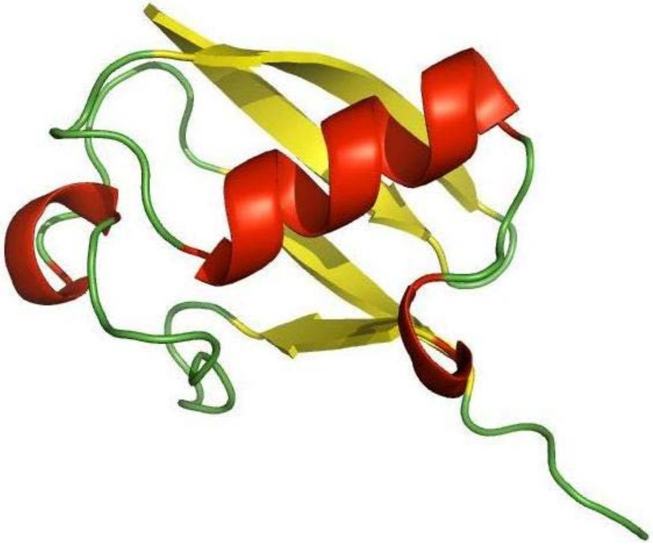
Example I:
Protein folding: The Levinthal paradox (1969)

- Assume: Two degrees of freedom for each residue(ψ,φ) with three values
- Possible conformations: 3^{2n}

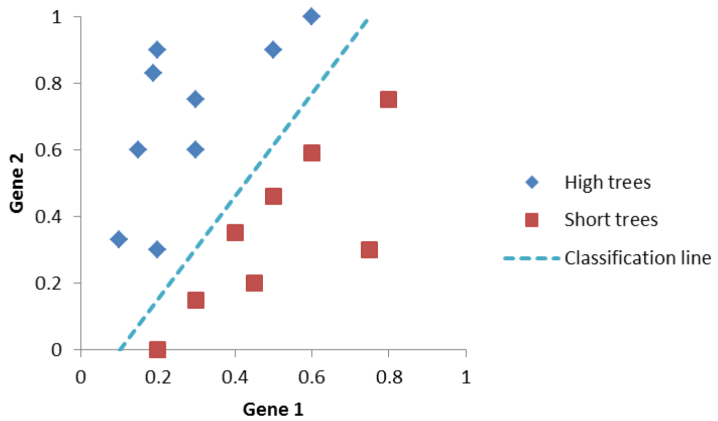
  (n is the number of amino acids)
- [$3^{2\cdot 84}$ particles in the universe]

- "Physics-based" protein folding is not realistic



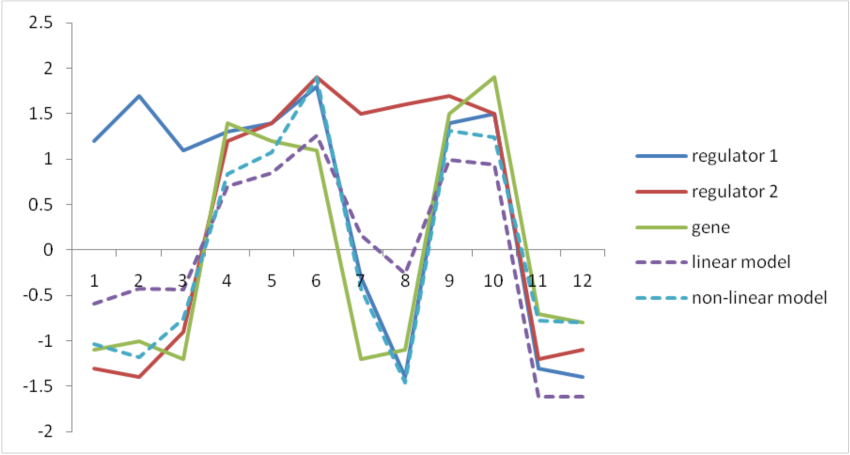


Example II:
Reductionism in genomics



Phenotypes emerge from interacting genes

Gene regulation include synergistic interaction such as AND-logic



Example III:
Inter-discipline reductionism

- Very few examples exists

- Thermodynamics could potentially be reduced to statistical mechanics
  - A shorthand for the language of statistical mechanics

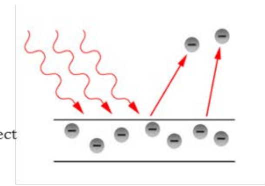
1. Temperature of a gas can be reducible to mean kinetic energy of the gas molecules
2. Gas pressure can be reducible to mean molecular density of the gas molecules

Gatherer, BMC Systems Biology 4:22, 2010

Is biology more than stamp collecting?

Are there general laws of molecular biology?

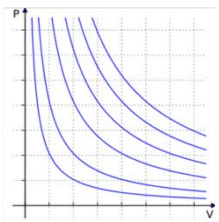
Physics
Photoelectric effect



$$E = h\nu$$

E = energy
H = Planck's constant
ν = frequency of light radiation

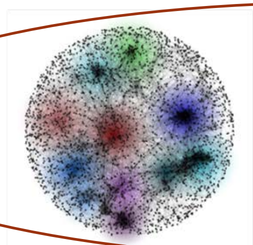
Chemistry
Ideal gas law



$$PV = nRT$$

P = absolute pressure
V = volume of the vessel
n = number of moles of gas
R = ideal gas constant
T = absolute temperature

Biology
Gene interactions



$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$

y_i = gene expression of gene i
n = number of genes
α = transcription rate
β_{ij} = effect of gene j on gene i

or is molecular biology a stamp collection?

1. …under heat stress gene A turn on gene B, that causes …
2. … the mutant show high levels of metabolite C indicating …
3. … five main gene expression clusters exists in developing cells of …
4. …

Biology's "theory of everything": Evolution

- Predicts how life started ("Biology's Bing Bang Theory")
  - A self-replicating molecule (RNA?)
  - that replicates with errors (mutations)

- John von Neumann's Universal Constructor (1940s): self-replicating machine in a cellular automata (CA) environment

- This seems like a good place to start looking for general laws …

… or is evolution the end of general laws?

- Evolution is a stochastic processes; isn't biology, like history, about describing what happened?

- Is evolution a scientific theory?
    - Homework: Can you design an experiment that, if successful, would falsify the theory of evolution?
- General laws in biology must be laws that evolved systems obey!

Molecular biology and the computer

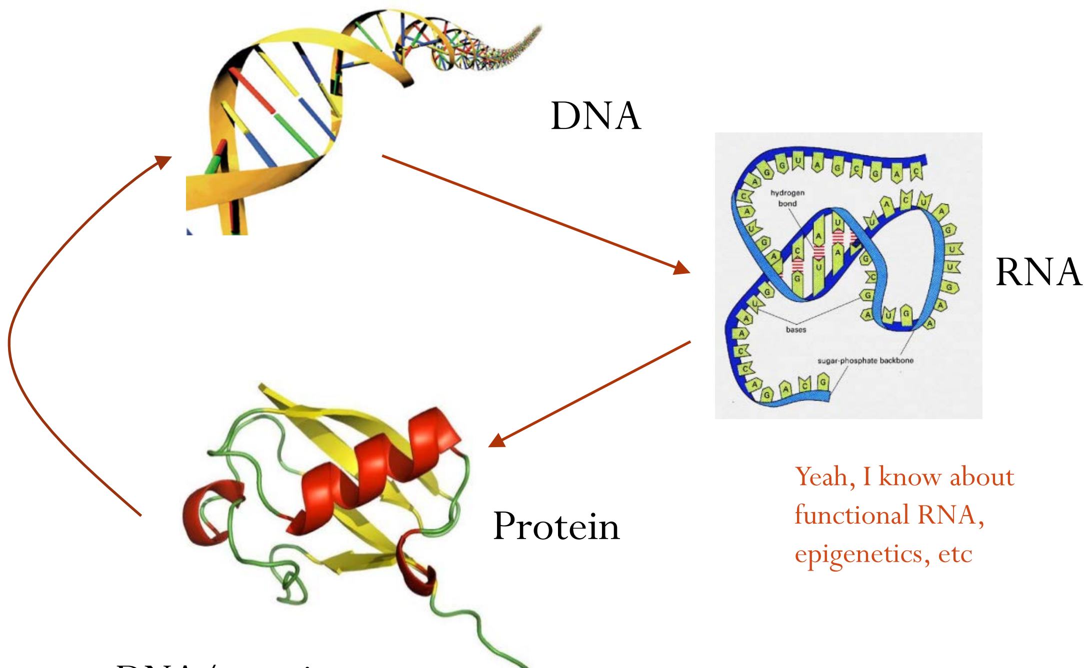
von Neumann (1945)

Memory

Buss

Central Processing Unit (CPU)

Evolution (4 000 000 000BC)

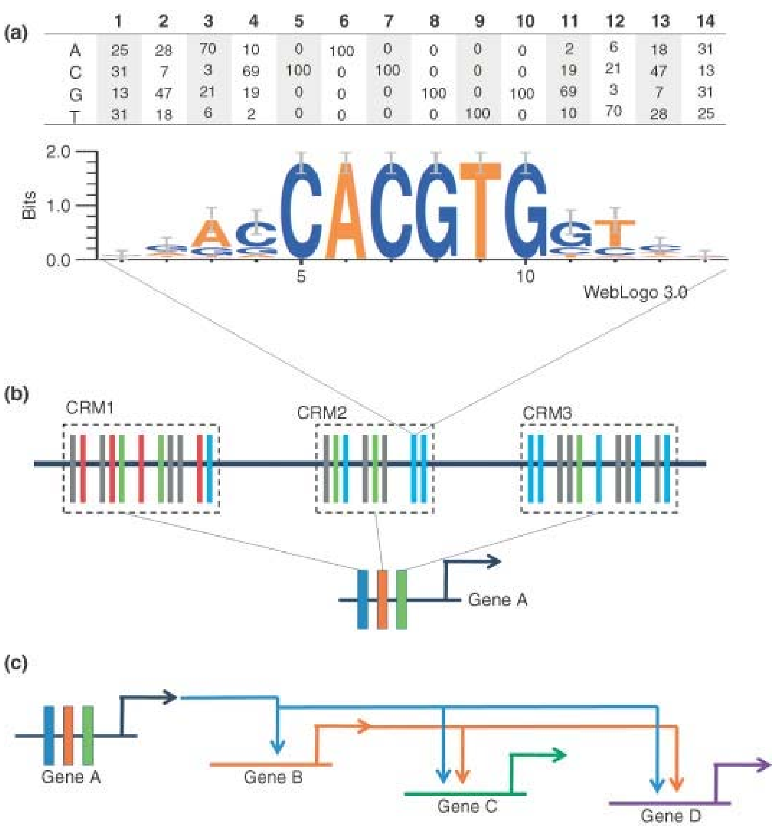


Yeah, I know about functional RNA, epigenetics, etc

Memory/CPU	→	DNA/protein
	→	Transcriptional network/Metabolic network
Instructions	→	Genes
Bus	→	RNA
Execution	→	Protein
Memory reference	→	Transcriptional control
Serial execution	→	Parallel execution

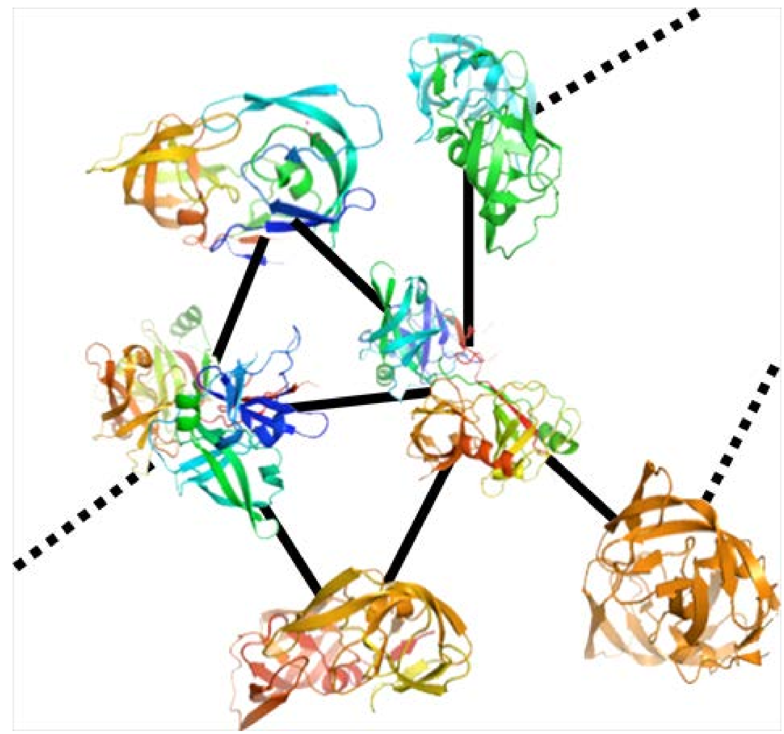
Molecular biology and hardware/software

Software → dynamic



The regulatory logics hard-wired in the regulatory genome

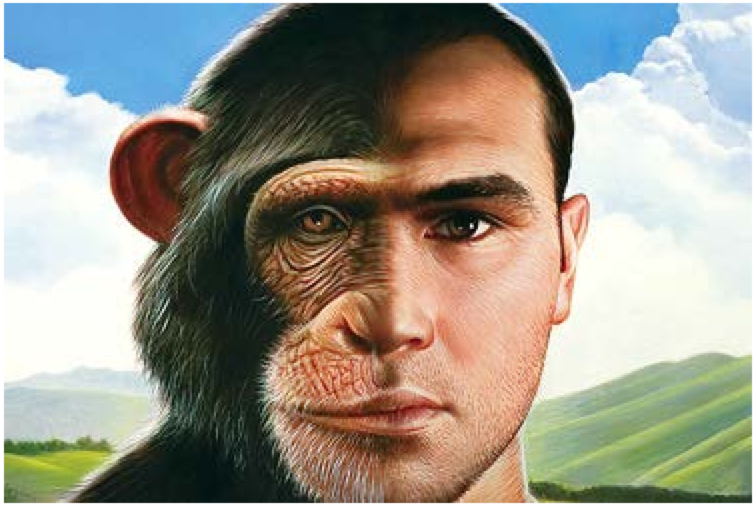
Hardware → static



The proteins are the building blocks of biological systems

Wilczynski *et al*. Fundamenta Informaticae 103: 323-332, 2010

Molecular biology and hardware/software



Humans and chimpanzee are 99% similar … in protein sequence — Hardware

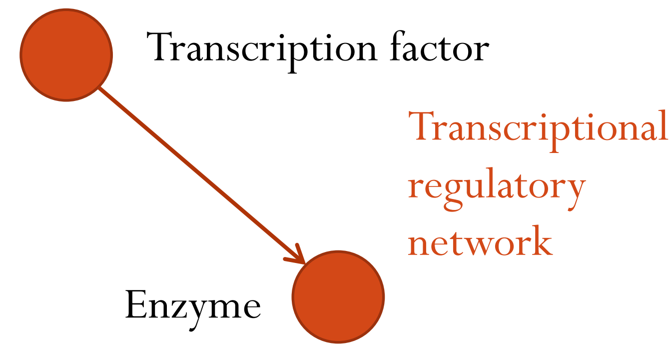
Differences are primarily in the regulatory genome — Software

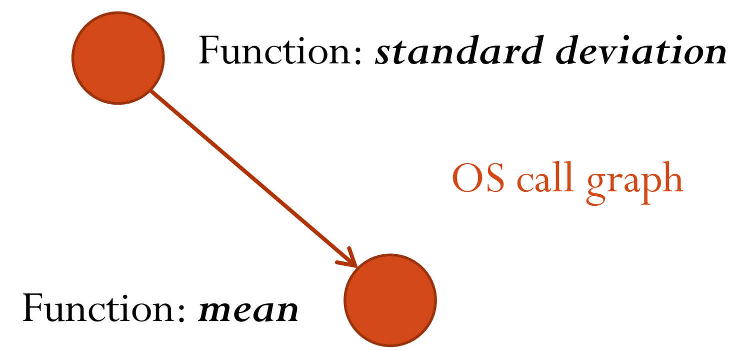
```
my @e1;
my @e2;
foreach my $s (keys %{$g{$t[$i]}}) {
      if (exists $g{$t[$j]}{$s}) {
            push @e1, $g{$t[$i]}{$s};
            push @e2, $g{$t[$j]}{$s};
      }
}
```

McLean *et al*. *Nature* 471: 216-219, 2011.

Comparing gene regulation and computer operating systems

- Operating system (OS): an interface between the hardware and application software
- An OS consists of thousands of functions that use each other
- Call graph: network visualizing how some functions use other functions





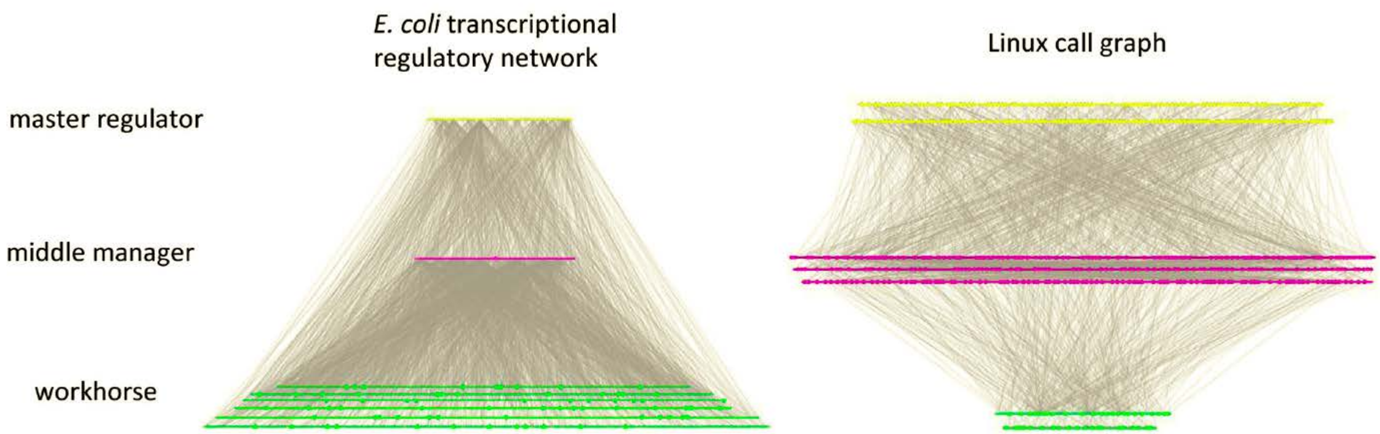
Yan *et al*. PNAS 107: 9186–9191, 2010.

Transcriptional regulatory network versus call graph

- Transcriptional regulatory network:
  - Organization: hierarchical, pyramid
- Call graph:
  - Organization: hierarchical, top-heavy
- Design: modules, persistent nodes

	E. coli transcriptional regulatory network	Linux call graph
Number of nodes	1,378	12,391
Number of persistent nodes	72* (5%)	5,120 (41%)
Number of edges	2,967	33,553
Number of modules	64	3,665
Number of comparative references	200 bacterial genomes	24 versions of kernels
Years of evolution	Billions	20

*In the *E. coli* genome 72 out of 212 persistent genes could be mapped to the transcriptional regulatory network.



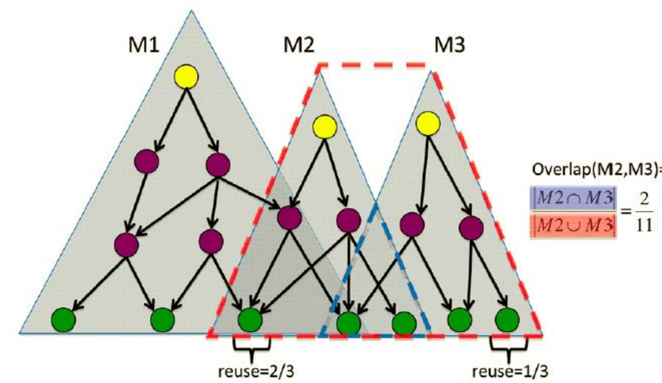
Yan *et al*. *PNAS* 107: 9186–9191, 2010.
Gama-Castro *et al*. *Nucleic Acids Res* 36:D120–124, 2008 (RegulonDB).

Transcriptional regulatory network versus call graph

- Transcriptional regulatory network:
  - Modules: non-overlapping (low reuse)
  - Persistent genes: workhorses (enzymes)
- Call graph:
  - Modules: overlap heavily (reuse)
  - Persistent functions: master regulators and middle managers

- Design principles: robustness in biology and cost effectiveness in software systems

A



B

	E. coli transcriptional regulatory network	Linux call graph
Number of modules	64	3665
Average overlap	4.3%	80.7%
Maximum node reuse	15.6%	87.5%
Average node reuse	3.5%	8.4%

Yan *et al*. PNAS 107: 9186–9191, 2010.

Laws of genome evolution

A. Log-normal distribution of the evolutionary rates between orthologous genes
B. Negative correlation between gene sequence evolution rate and expression level (or protein abundance)
C. Power law–like distributions of membership in paralogous gene families and node degree in biological networks
D. Distinct scaling of functional classes of genes with genome size


OPEN ACCESS Freely available online

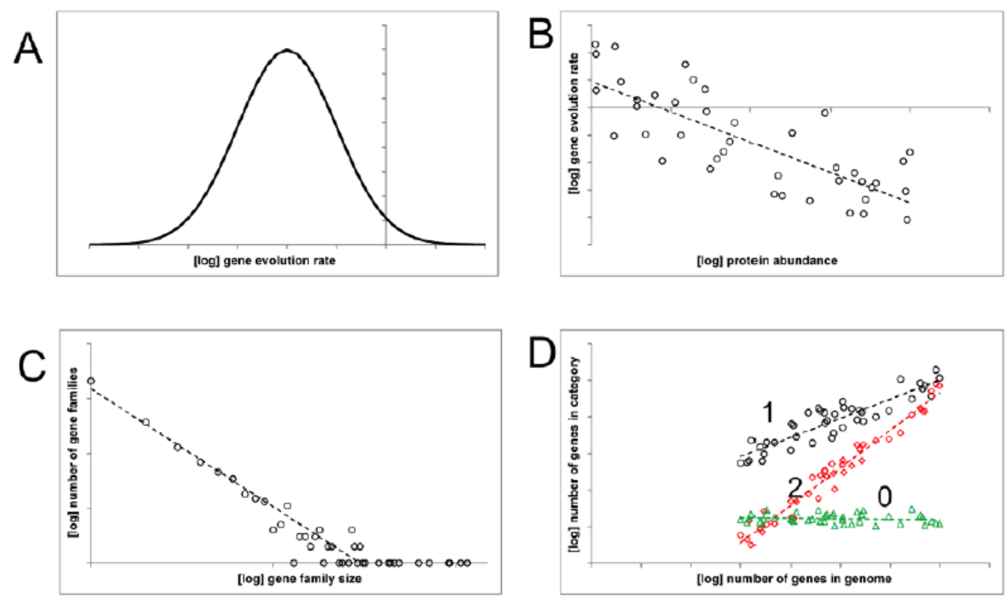
PLoS COMPUTATIONAL BIOLOGY

Review

Are There Laws of Genome Evolution?

Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America



0. No dependence: translation
1. Linear dependence: enzymes
2. Quadratic dependence: regulation/signaling

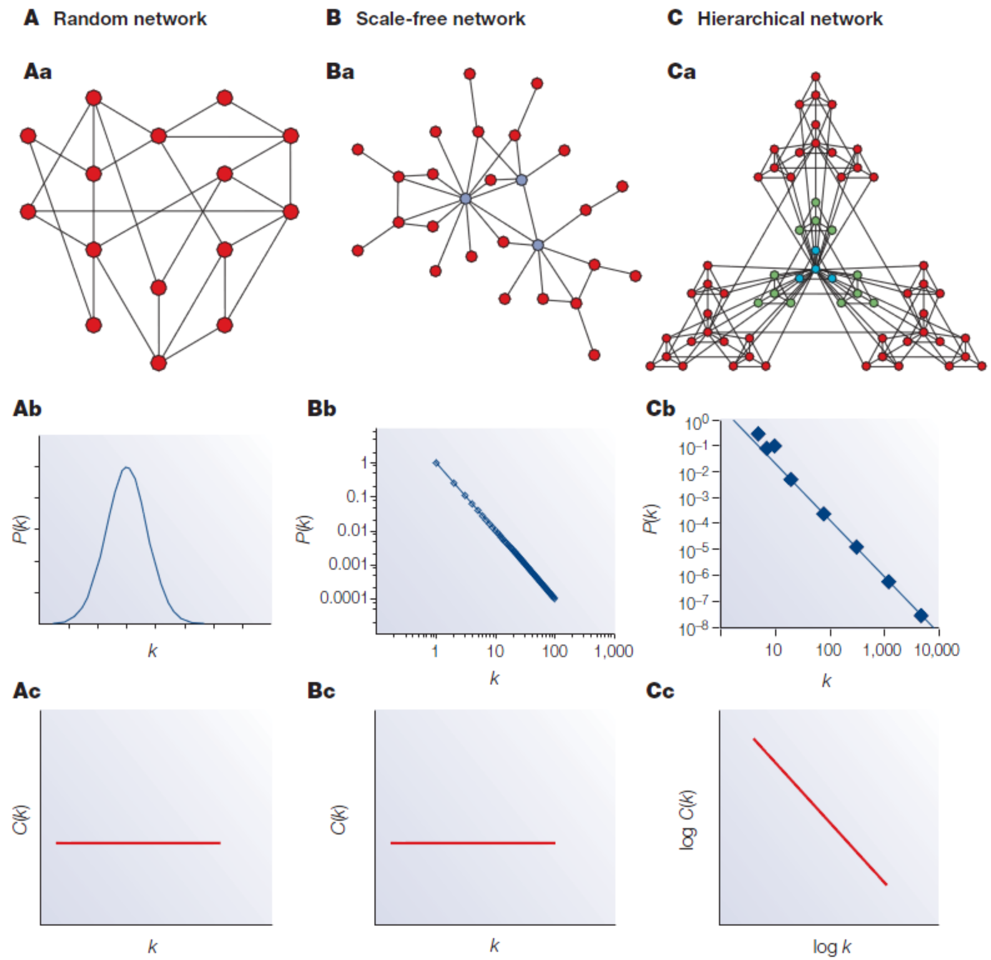
Koonin. PLoS Computational Biology 7:e1002173, 2011.

Power law-like node degree distribution in biological networks

- *k* – node degree: the number of links a node has to other nodes
- *P(k)* – the degree distribution
- Scale-free network:

  $P(k) \sim k^{-\gamma}$, where γ is the degree exponent ($2 < \gamma < 3$)

- *C*(*k*) – clustering coefficient (tendency of nodes to form clusters)

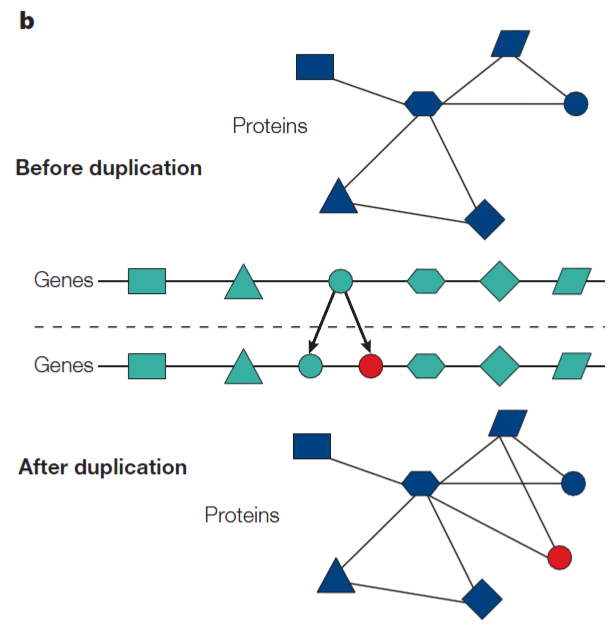


Barabási and Oltvai. *Nature reviews* 5: 101-113, 2004.

Scale-free networks: Does it mean anything?

- Scale-free networks are robust to random error, but vulnerable to attack
- Gene evolution by "duplication followed by subfunctionalization" also explains the scale-freeness of networks
- Robustness may emerge "for free"

- Scale-freeness is an emergent property that appear because networks consist of numerous genes/proteins that weakly interact with each other
  - Analogous to ideal gases described by statistical mechanics



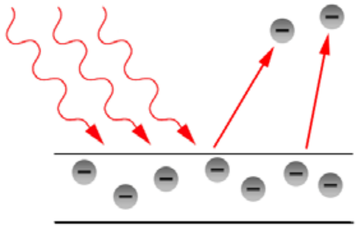
Barton and Coe. On the application of statistical physics to evolutionary biology. *Journal of Theoretical Biology* 259: 317–324, 2009.

Barabási and Oltvai. *Nature reviews* 5: 101-113, 2004

Koonin. *PLoS Computational Biology* 7:e1002173, 2011

Physics

Photoelectric effect

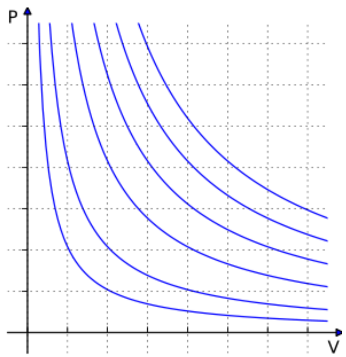


$$E = h\nu$$

E = energy
H = Planck's constant
ν = frequency of light radiation

Chemistry

Ideal gas law



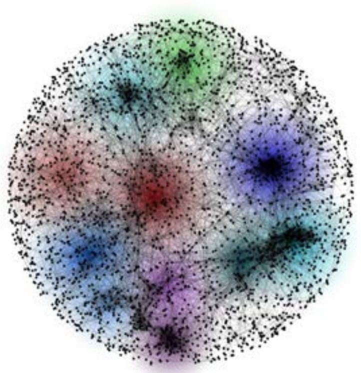
$$PV = nRT$$

P = absolute pressure
V = volume of the vessel
n = number of moles of gas
R = ideal gas constant
T = absolute temperature

Biology

~~Gene interactions~~

Scale free networks



~~$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$~~

~~y_i = gene expression of gene i~~
~~n = number of genes~~
~~α = transcription rate~~
~~β_{ij} = effect of gene j on gene i~~

$$P(k) \sim k^{-\gamma}$$

k = node degree
P(k) = degree distribution
γ = degree exponent

Anyway, here is my five cents: Is biology more than stamp collecting?

For

- There is a theory of everything: evolution
- General design principle exists in evolved systems
- Evolved systems do obey general laws (e.g. power law)
- More and better data will give better models
- Models may be more complex than in physics, but at some point biology will be "understood"

Against

- Many biologists don't care all that much about general laws
- Biologists are more interested in exceptions than rules
- Inter-discipline reductionism is too hard; biology is too complex
- Evolution has given us a history not the history

Thank you for listening!

Bonus slides

Occam's razor

- William of Occam 14th century: *things should not be multiplied unnecessarily*
- Issac Newton (1687): *we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearance*
- Albert Einstein (20th century): *everything should be made as simple as possible, but not simpler*
- *The simplest model that explains the data should be chosen*
  - *Require weighting model complexity (no. parameters) against model fit (p-value)*
  - *Example: multiple hypothesis testing (significance threshold = 0.05 / n)*

My interpretation of Rutherford's statement "All science is either physics or stamp collecting"

- A belief that science is all about finding general laws
- A belief in inter-discipline reductionism

- A sign of the high self confidence of physicists at the time
- A sign of their (over-)confidence in mathematics

- Ironically, Rutherford got a Nobel Prize in Chemistry (1908)

More relevant quotes:

- “The exception proves that the rule is wrong. That is the principle of science. If there is an exception to any rule, and if it can be proved by observation, that rule is wrong”. Richard Feynman.
- “The great tragedy of Science — the slaying of a beautiful hypothesis by an ugly fact”. Thomas Henry Huxley
- “All models are wrong, but some are useful”. George E. P. Box
- “Science may be described as the art of systematic oversimplification.” Karl Popper
- “Physics is like sex: Sure, it may have practical results, but that is not the reason we do it”. Richard Feynman
- “Prediction is very difficult, especially about the future”. Niels Bohr
- “Today's scientists have substituted mathematics for experiments, and they wander off through equation after equation, and eventually build a structure which has no relation to reality”. Nikola Tesla
- “Computer Science is no more about computers than astronomy is about telescopes”. Edsger Dijkstra
- “Type III error: finding the right answer to the wrong question”. A. W. Kimball