Human viewed as a machine

or Biology from the point of view of a Mathematician

Alvis Brazma

Bioinformatics Summer School, BDS 3, Uzhhorod, July 2023

EMBL-EBI

The world before 1991

Riga

With two of my fellow undergraduate students in the Carpathian mountains not too far from here in ~1978

ChatGPT, continue:

continue 16, 32, 64, 128, 256, ...

512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 262144, ...

continue 16, 1, 6, 32, 3, 2, 64, 6, 4, 128, 1, 2, 8, 256, 2, 5, 6, ...

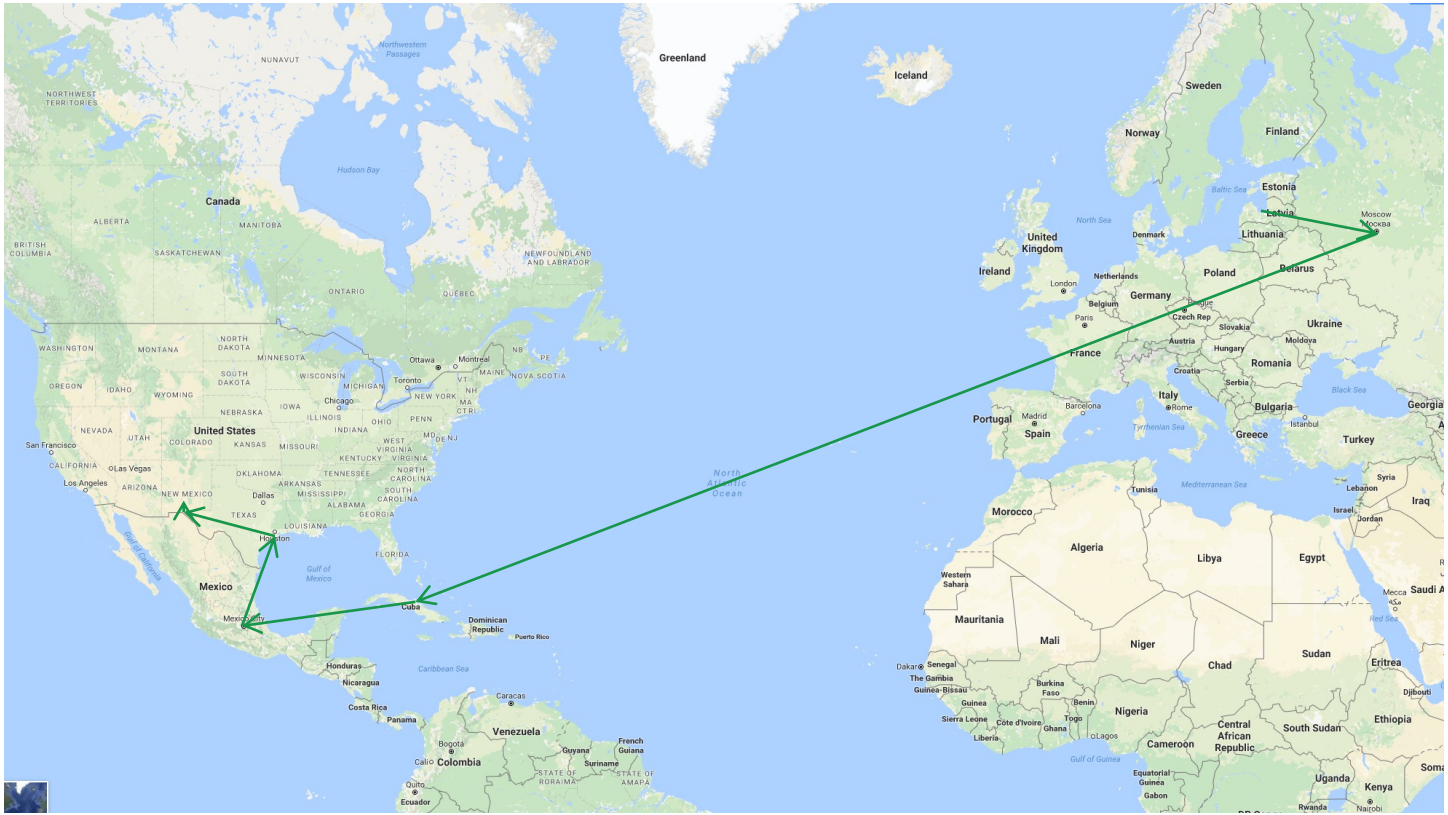
This sequence doesn't seem to have an obvious pattern, but here are the next few terms:

512, 4, 8, 1024, 9, 10, 2048, 12, 16, 4096, 1, 18, 32, 8192, 20, 25, 18, 16384, 27, 4, 65536, 30, 32, 512, ...

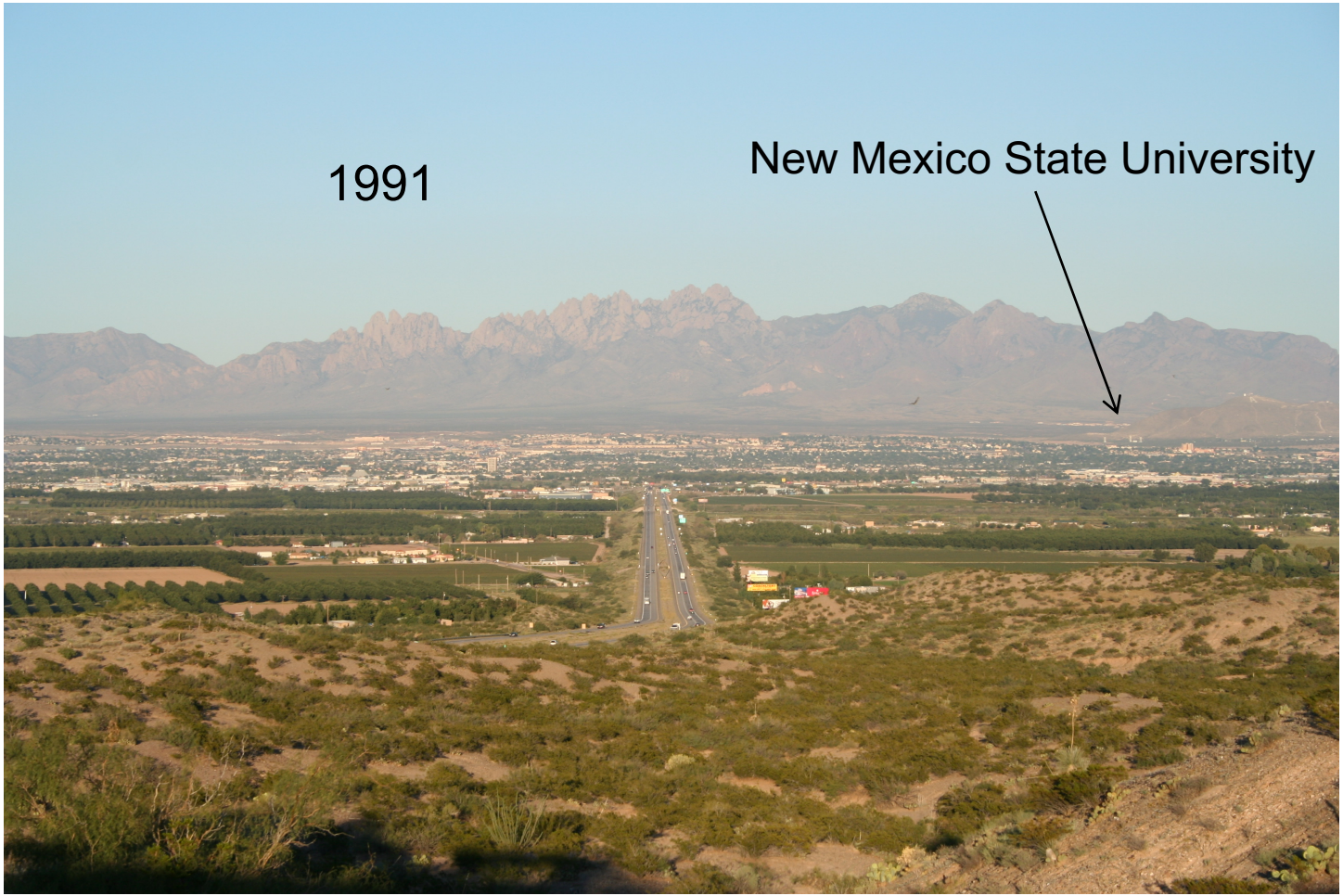
Without more information, it's difficult to determine a clear rule or formula that generates this sequence. It's possible that the sequence is arbitrary or follows a complex pattern that is not immediately apparent.

My PhD was roughly about this

9 November 1989



My itinerary on 14-16 January 1990


1991
New Mexico State University

The case for the human genome

A meeting next month in London should give a clear view of where the Human Genome Project is now, and where it is likely to be heading in the years that follow. The project (which is several) continues to gather momentum. The strategic questions have been settled, but there will be endless arguments on tactics. And ethical issues, while insubstantial, must be handled delicately.

FAMILIARITY exorcizes fear. That seems to be one of the unremarkable lessons learned so far from the great Human Genome Project or, more accurately, from the dozens of national projects that have sprung up in the past few years in different places.

Fears that research in biology would in future be conducted on automated assembly lines, or that research laboratories would be robbed by the demands of the sequencing plants of a whole generation of skilled technicians, have not materialized and are unlikely to do so, perhaps because none of the projects has yet entered what might be called its production phase.

Yet on the first issue, it is curious that Professor Walter Gilbert's impassioned statement of the reasons why biologists must learn

plete sequence began.

Fears that human genome projects everywhere would rob other worthwhile projects of funds have similarly not yet bitten into the pattern of research supported by the grant-making agencies. Moreover, when people are talking realistically of automatic sequencing at a cost of no more than $0.50 a base (with overheads included), it seems unlikely that a ten-year project to provide the complete sequence of 23 stretches of DNA, with 3,000 million bases altogether, should much exceed the annual cost of the present development phase.

The ethical objections to the sequencing of the human genome are necessarily more subtle. They range from the assertion that it would be improper that knowledge won in a

than any now available to the places in the human genome at which genetic abnormalities occasion genetic disease or a high frequency of somatic genetic change (as in cancer) — that is one of the important objectives of the exercise, after all. But the objection, if valid, also applies to what is already possible by the use of genetic markers for identifying particular variants of a genetic polymorphism.

In present practice in, for example, the insurance of persons, the trouble (and cost) that companies take to discriminate between their would-be customers is measured by the risks they are asked to undertake; it remains to be seen whether present ambitions to automate the sequencing process will be followed by a machine that can print out a

European Bioinformatics Institute

- World leading source of public genomics and biomolecular data
- We accept, process, archive and make available to the life sciences community genomics and other life sciences data
- We are part of the European Molecular Biology Laboratory (EMBL), Europe's flagship laboratory for the life sciences.

EMBL-EBI

EMBL member states

Member states (29)

Austria 1974
Denmark 1974
France 1974
Germany 1974
Israel 1974
Italy 1974
Netherlands 1974
Sweden 1974
Switzerland 1974
United Kingdom 1974
Finland 1984
Greece 1984
Norway 1985
Spain 1986
Belgium 1990
Portugal 1998
Ireland 2003
Iceland 2005
Croatia 2006
Luxembourg 2007
Czech Republic 2014
Malta 2016
Hungary 2017
Slovakia 2018
Montenegro 2018
Poland 2019
Lithuania 2019
Estonia 2023
Latvia 2023

Associate member states

Australia 2008

Prospect member states

Serbia

EMBL-EBI

European Molecular Biology Laboratory - EMBL

Europe's centre of excellence in *life sciences* research, services and training

Founded in 1974 by 10 states as an intergovernmental organisation to promote the molecular life sciences in Europe and beyond

"I believe that international activity is very important in building world peace." Sir John Kendrew, EMBL's 1st DG

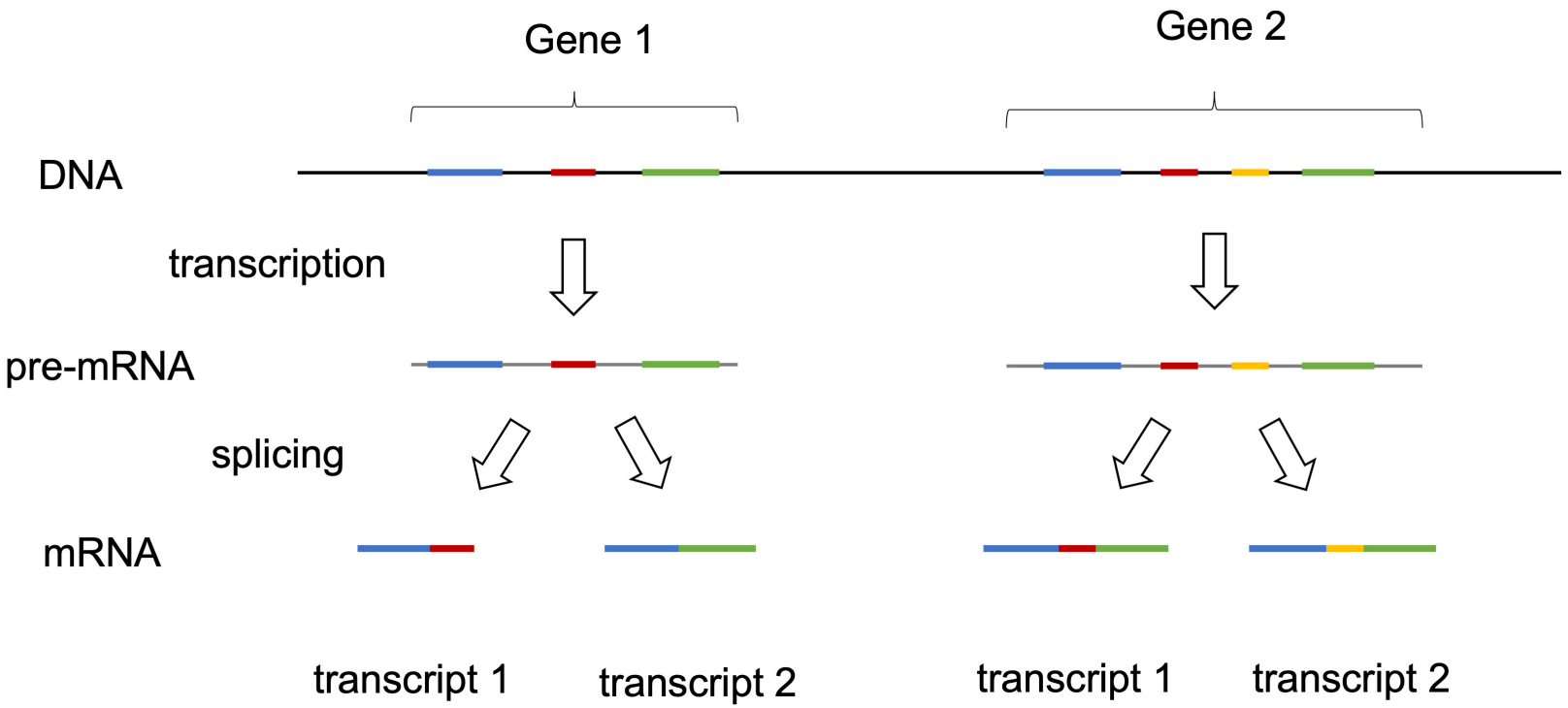
First Director General
Sir John Kendrew
Nobel Prize in 1962

EMBL-EBI

What's the difference between theoretical and experimental biologists?

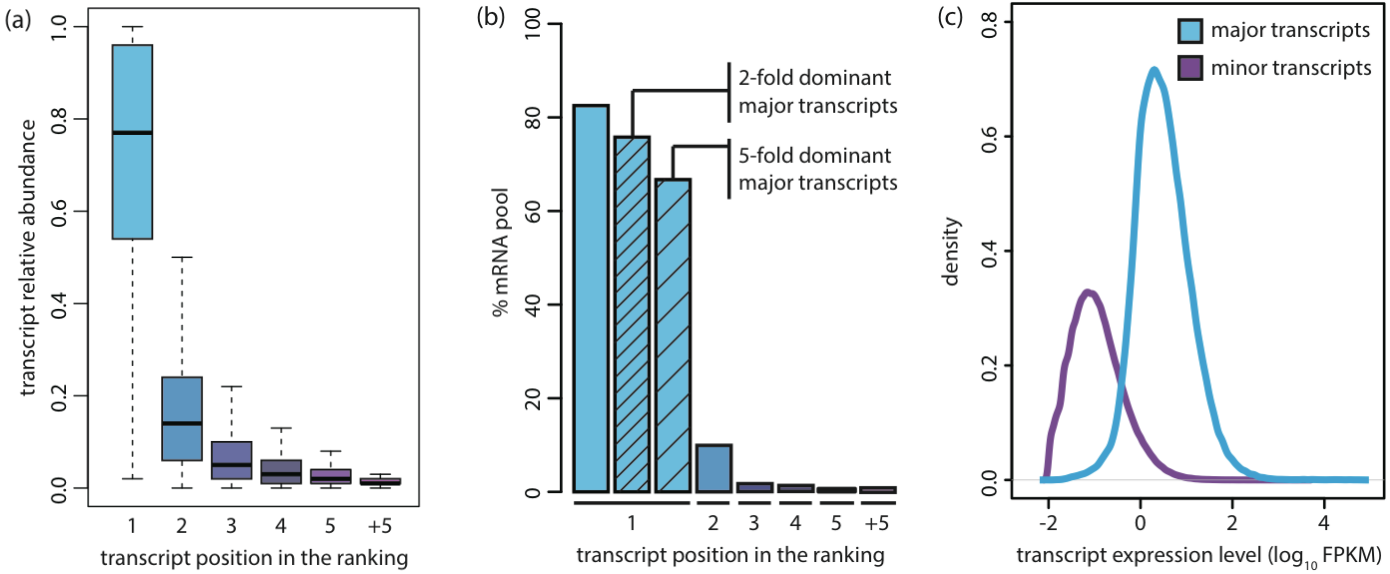
- Experimental biologists observe what cannot be explained
- Theoretical biologists explain what cannot be observed

- Two perspectives on science
  - Mathematicians – looking for rules from which everything else follows as much as possible; exceptions can be taken care of later
  - Biologists – outliers (exceptions) are what is interesting, the rules, if there are any, usually are obvious and not interesting


Gene 1
Gene 2
DNA
transcription
pre-mRNA
splicing
mRNA
transcript 1
transcript 2
transcript 1
transcript 2

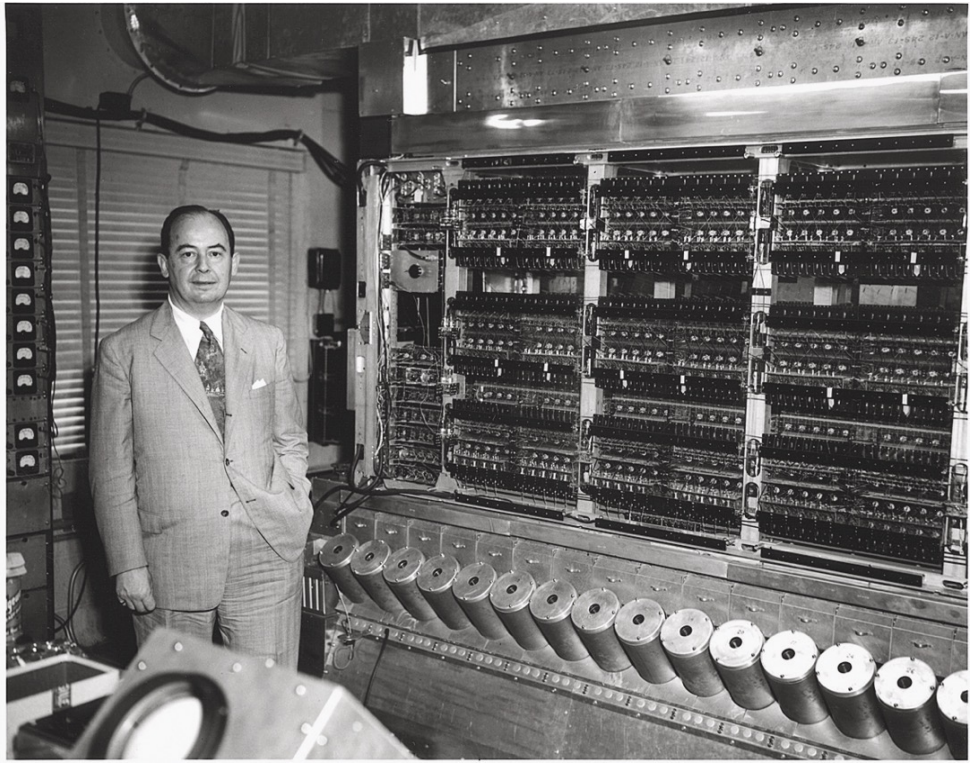
An example of a rule (with many important exceptions) – most genes in most tissues express only one splice-form

Mar Gonzàlez-Porta



EMBL-EBI

Can a machine reproduce? What are the minimum needed for this?



John von Neumann (1903 – 1957)

Ability to reproduce is often seen as one of the dividing lines between living and nonliving

Her Majesty pointed to a clock and said, 'See to it that it produces offspring'.

From alleged conversation between Queen Christina of Sweden and René Descartes in 1649

Von Neuman's mechanical replicator

- *Draw up a list of unambiguously defined elementary parts. Imagine that there is a practically unlimited supply of these parts floating around in a large container. One can imagine an automaton functioning in the following manner: It also is floating around in this medium; its essential activity is to pick up parts and put them together (…).* Von Neumann and Burks (1966) from von Neumann's lectures in 1948

- The task of this automaton is to assemble a clone of itself

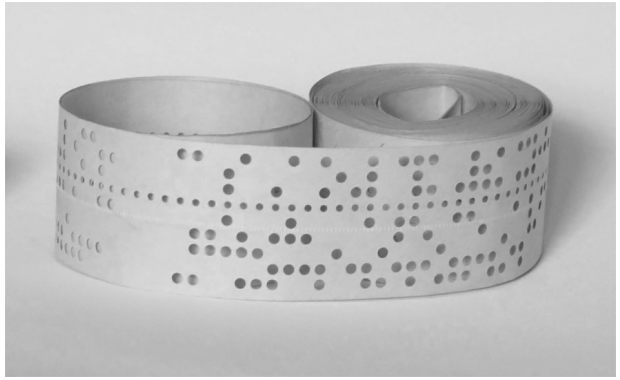
Degenerative trend

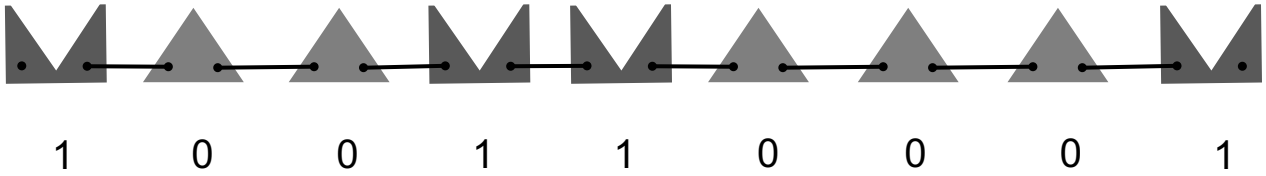
- It does appear that for an *automaton* to make *a thing*, the *automaton* needs to contain the full description of that *thing*
- If this description is a part of the *automaton*, then the automaton becomes more complex than the *thing* it can build
- Doesn’t it follow that an *automaton* can only make *things* that are simpler that itself? If so, this leads to a *degenerative trend.* Isn’t this a paradox?

- The solution to this paradox is in the decoupling the description (of the *thing* to build) from the *automaton* building it

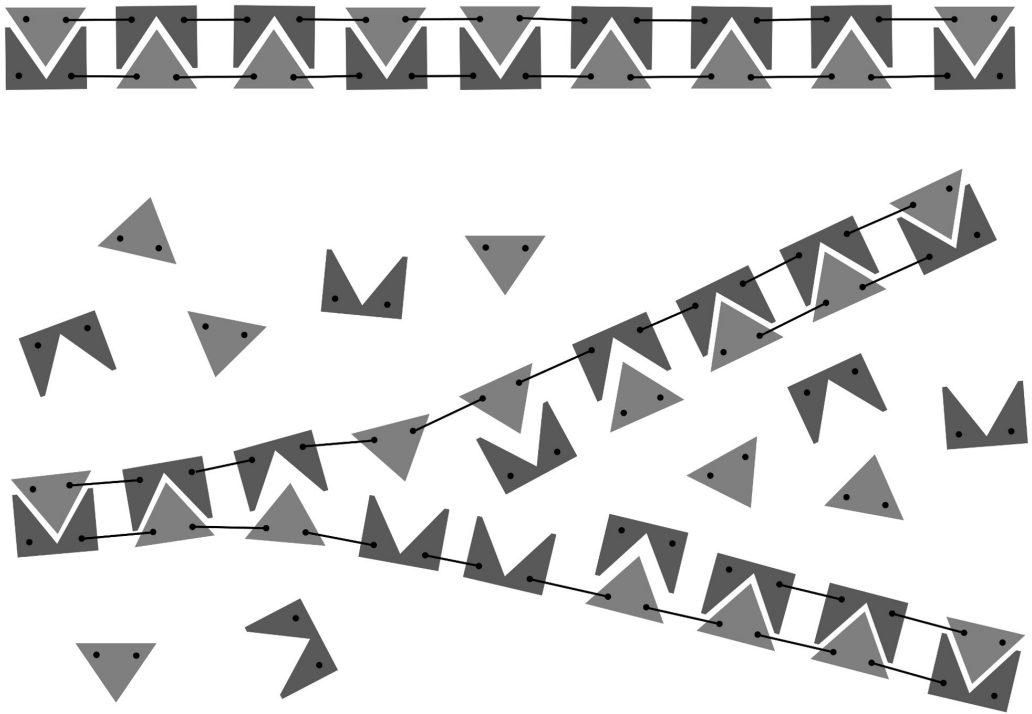
The Universal Constructor

- *Universal Constructor*—an abstraction of a general-purpose robot, which given appropriate instructions and necessary materials could build whatever it has been instructed to build.
- Instructions to the universal constructor can be encoded on a **tape**
- Replicating 1- or 2-dimensional objects are much simpler that replicating a complex 3-diminensional object





Exploiting complementarity



How to clone oneself

- X – an arbitrary *thing* made of von Neumann's elementary parts; this *thing* can also be some sort of *aparatus*
- U – Universal Constructor
- Code(X) – a tape containing the description of X that given to U produces X
- R – a simple device that can replicate the tape Code(X)
- C – a *controlling device* that given Code(X) first runts R on Code(X) producing another copy of Code(X), then runs U on Code(X) producing X and then 'ties' them together

How to clone oneself

- The device U+R+C+code(X) produces X+code(X)

- Thus, starting with U+R+C+code(X) after some time we get

  U+R+C+code(X), X+code(X)

- But X is an arbitrary object, it can be some sort of *apparatus*. Take

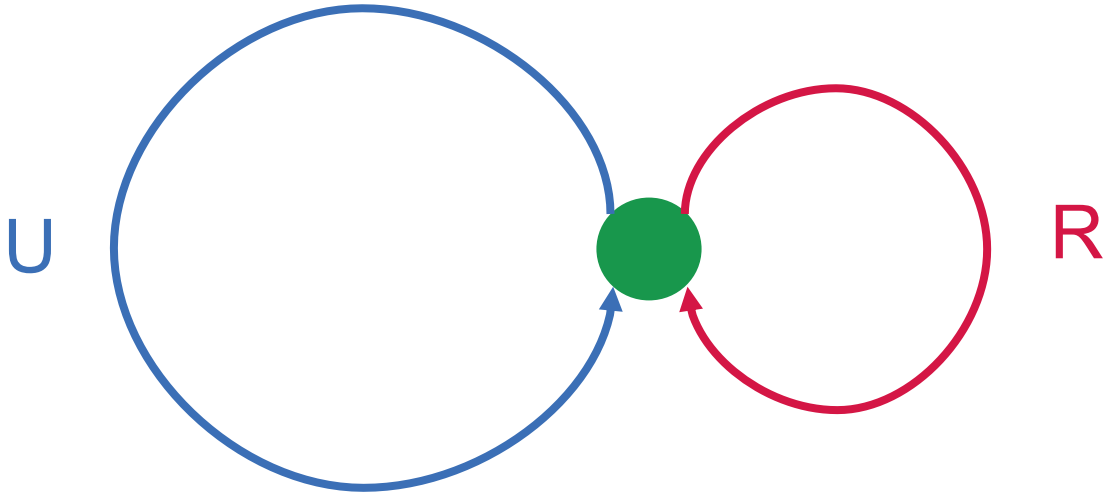
  X = U+R+C+code(U+R+C)

- Then starting from U+R+C+code(U+R+C) we get

  U+R+C+code(U+R+C), U+R+C+code(U+R+C)

The two processes U and R have to be separated

To beat the *degenerative trend* processes U and R have to be separated



Crick's central Dogma

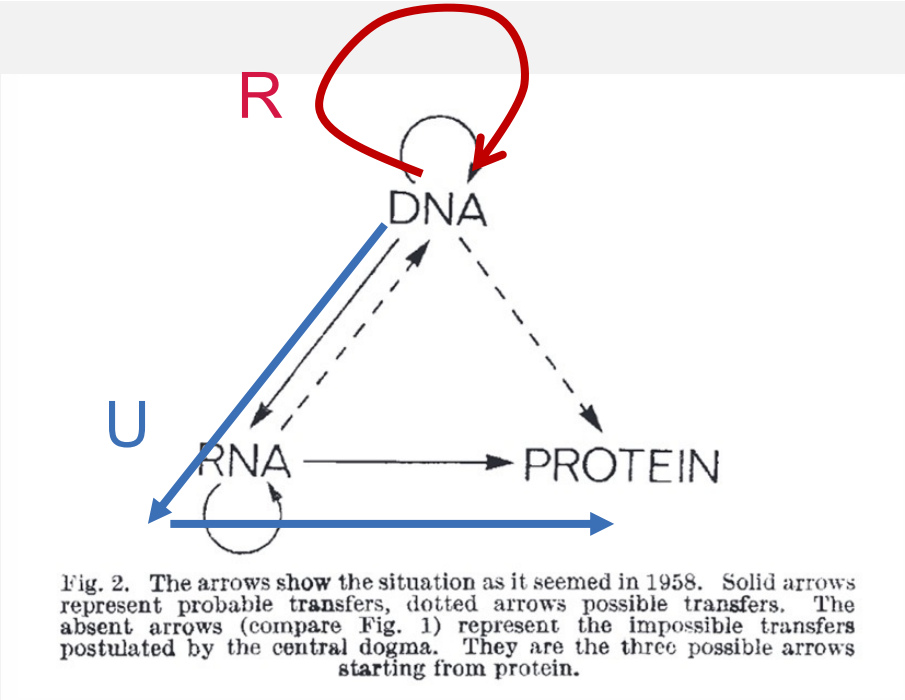
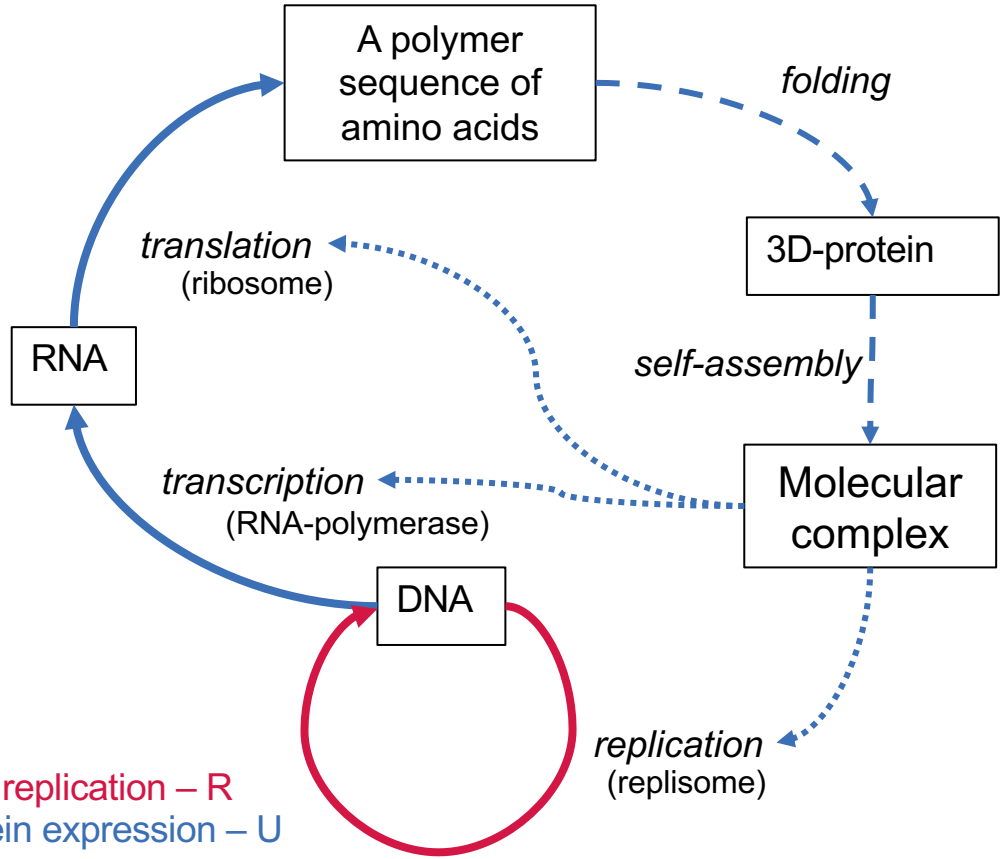
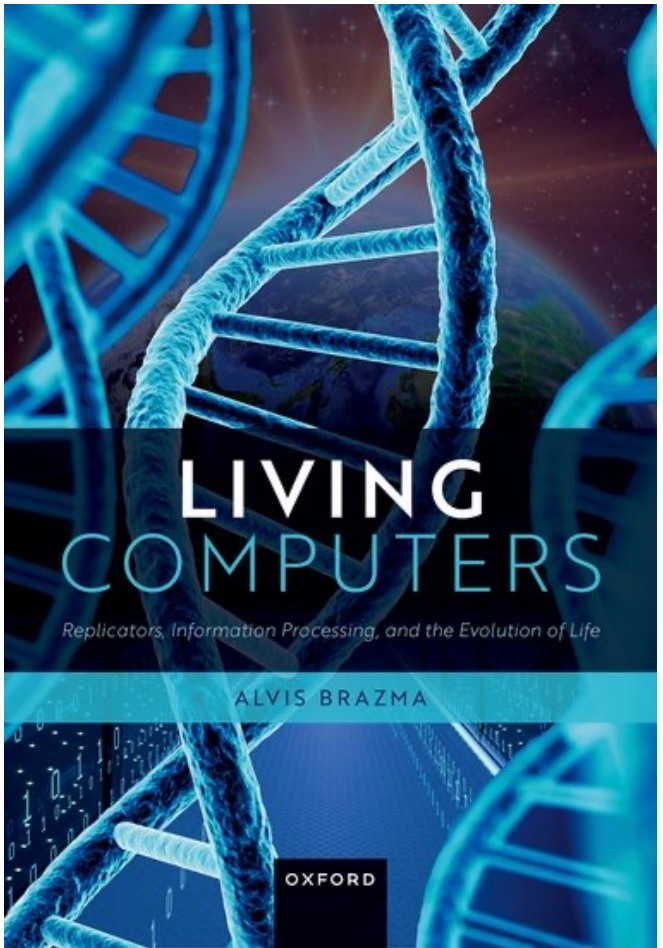


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

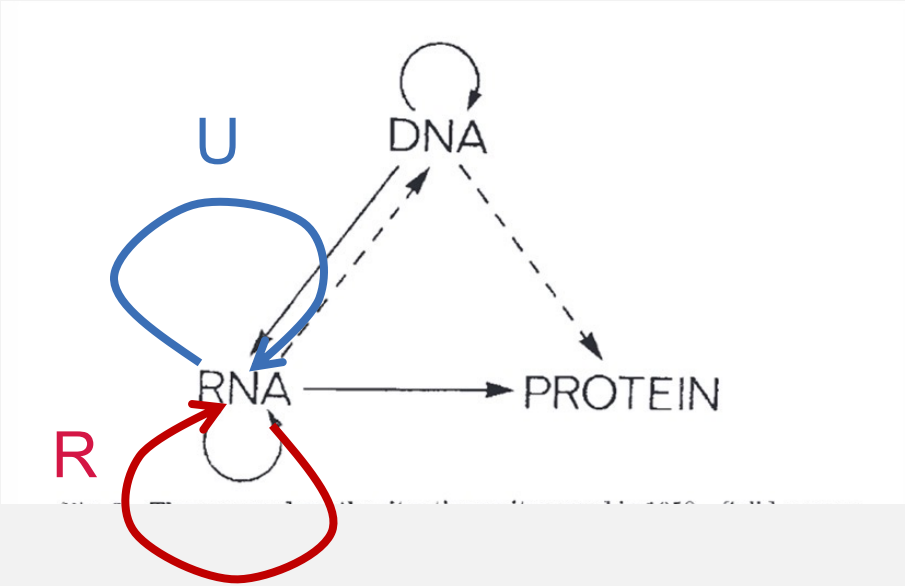
Crick's central dogma and von Neuman's replicator



DNA replication – R
Protein expression – U
Self-organisation – C



RNA life



Is this how life started?

Minimum complexity threshold

- There is minimum complexity necessary for information carrying self-replication.
- Below this complexity we have only degenerative trend possible
- Above this the system can build a machine as as complex as itself and (through evolution) even a more complex system

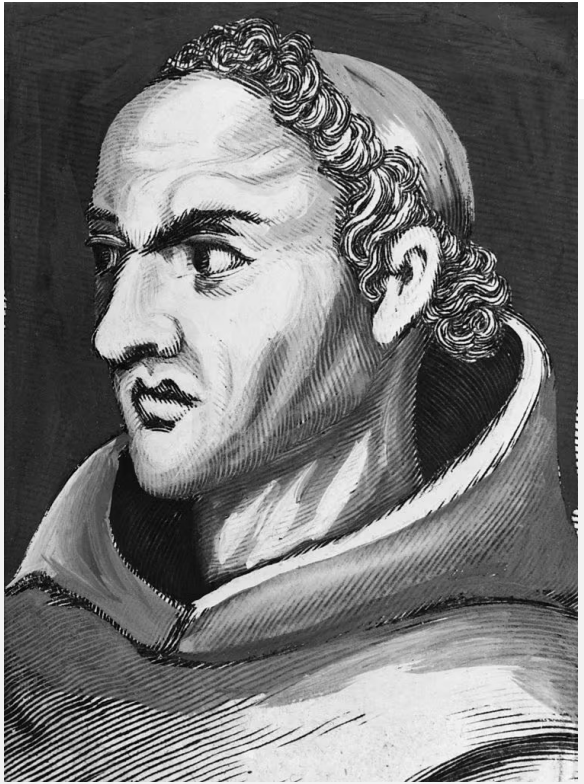
- Given that Universal Constructor is quite complex, this complexity threshold is quite high

- Life cannot be too simple!
- Biologists will say ”we knew it all along”. For a mathematician this is a revelation ;-)

Occam's Razor

- **"Entities should not be multiplied without a necessity"**
- The modern formulation – *if several theories explain the observations equally well, choose the simplest one*

- In my own formulation – *explain as much as possible with as little as possible*



William of Ockham (1287 – 1347)

Rissanen's minimum description length (MDL) principle:

the best model of data is their shortest description (in bits)

This is a practical implementation of Occam's Razor

Applying Rissanen's MDL principle

```
ATTGATGAGAGTTA
TAGCAGGATGAGTAGCAG
TTAGCAGGATGAGCTGCAT
GAGCTGATGAGTTACA
CAGTCTGATGAGGCGTATA
TATTGTGAAATATTTTATTG
CCATTGAAACGCAGCATGAT
CTAGTTGTGAAAGCGAGCTGATG
ATCTCGTGAAAGGTATTTGCAG
TGATGAAACGAAATGAAA
```

Applying Rissanen's MDL principle

ATTGATGAGAGTTA
TAGCAGGATGAGTAGCAG
TTAGCAGGATGAGCTGCAT
GAGCTGATGAGTTACA
CAGTCTGATGAGGCGTATA
TATTGTGAAATATTTTATTG
CCATTGAAACGCAGCATGAT
CTAGTTGTGAAAGCGAGCTGATG
ATCTCGTGAAAGGTATTTGCAG
TGATGAAACGAAATGAAA

Applying Rissanen's MDL principle

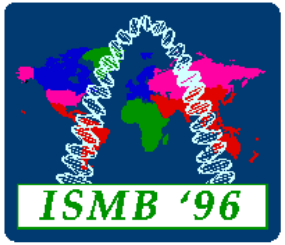
```
ATTGATGAGAGTTA
TAGCAGGATGAGTAGCAG
TTAGCAGGATGAGCTGCAT
GAGCTGATGAGTTACA
CAGTCTGATGAGGCGTATA
TATTGTGAAATATTTTATTG
CCATTGAAACGCAGCATGAT
CTAGTTGTGAAAGCGAGCTGATG
ATCTCGTGAAAGGTATTTGCAG
TGATGAAACGAAATGAAA
```

189 letters

```
ATTX1AGTTA
TAGCAGX1TAGCAG
TTAGCAGX1CTGCAT
GAGCTX1TTACA
CAGTCTX1GCGTATA
TATTGX2TATTTTATTG
CCATX2CGCAGCATGAT
CTAGTTGX2GCGAGCTGATG
ATCTCGX2GGTATTTGCAG
TGAX2CGAAATGAAA

X1=GATGAG
X2=TGAAA
```

171 characters



Discovering Patterns and Subfamilies in Biosequences

Alvis Brāzma*
abrazma@cclu.lv
Institute of Mathematics and Computer Science
University of Latvia
29 Rainis Bulevard
LV-1459 Riga, Latvia

Inge Jonassen
inge@ii.uib.no
Department of Informatics
University of Bergen, HIB
N5020 Bergen, Norway

Esko Ukkonen
Esko.Ukkonen@cs.Helsinki.FI
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Jaak Vilo
Jaak.Vilo@cs.Helsinki.FI
Department of Computer Science
P.O.Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki
Finland

Abstract

We consider the problem of automatic discovery of patterns and the corresponding subfamilies in a set of biosequences. The sequences are unaligned and may contain noise of unknown level. The patterns are of the type used in PROSITE database. In our approach we discover patterns and the respective subfamilies simultaneously. We develop a theoretically substantiated significance measure for a set of such patterns and an algorithm approximating the best pattern set and the subfamilies. The approach is based on the minimum

2. as grouping or clustering a set of biosequences into subsets so that each subset shares a distinct common pattern and the noise is sorted out.

A subproblem of this is finding a pattern common to a set of sequences in the presence of an unknown level of noise.

In particular we consider sequences representing proteins and patterns of the type used in PROSITE database (Bairoch 1992), but the method can be applied

$$F'(\pi, l) = l \cdot (I(\pi) + 2\log p + f(\pi)) - M_1(\pi),$$

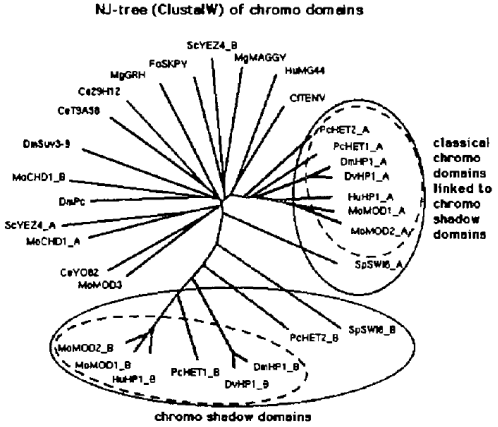
where

$$I(\pi) = I_1(\pi) + I_2(\pi),$$

$$I_1(\pi) = \sum_{j \in \dot{\Phi}(\pi)} I'(\pi(j)),$$

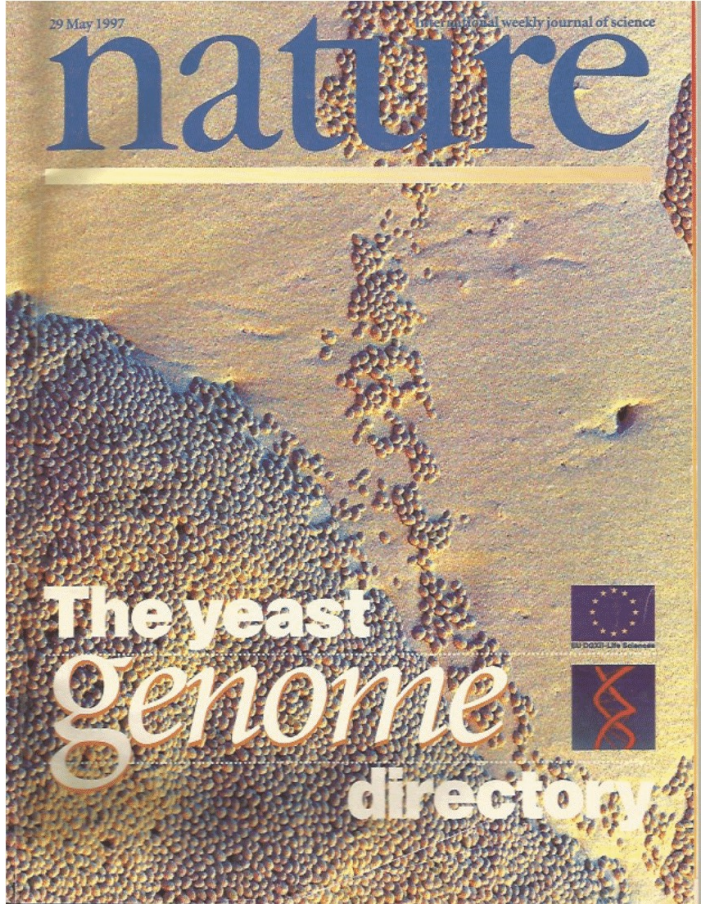
$$I'(b_h) = -\sum_{i=1}^{m} P_2(a_i) \log P_2(a_i) + \\ + \sum_{a_i \in K_h} P_2(a_i|b_h) \log P_2(a_i|b_h),$$

$$I_2(\pi) = -\sum_{x(t_i, v_i)} \log(v_i - t_i + 1)$$

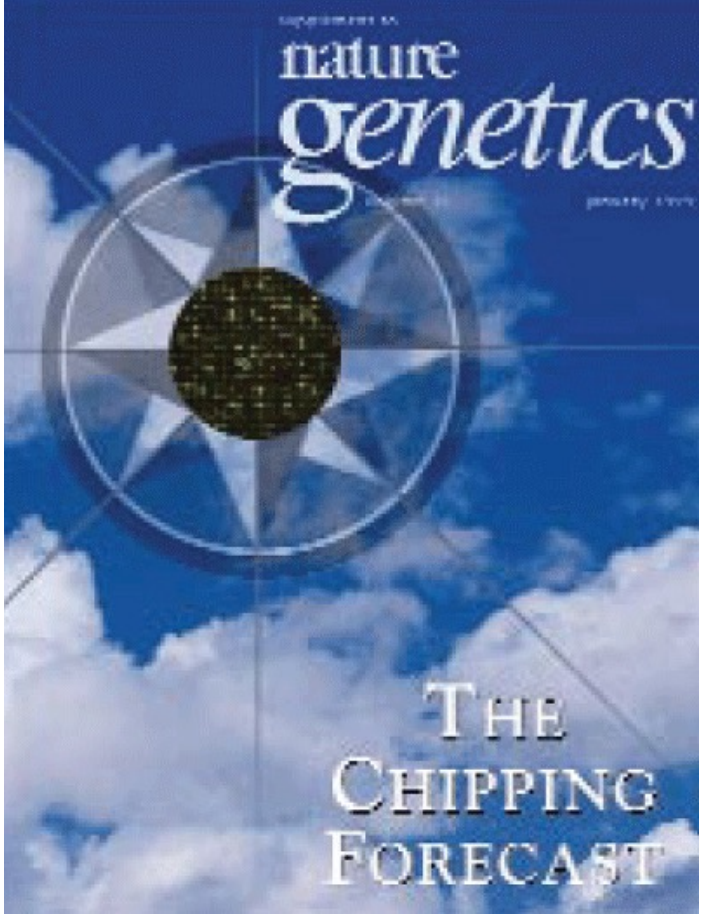


EMBL-EBI

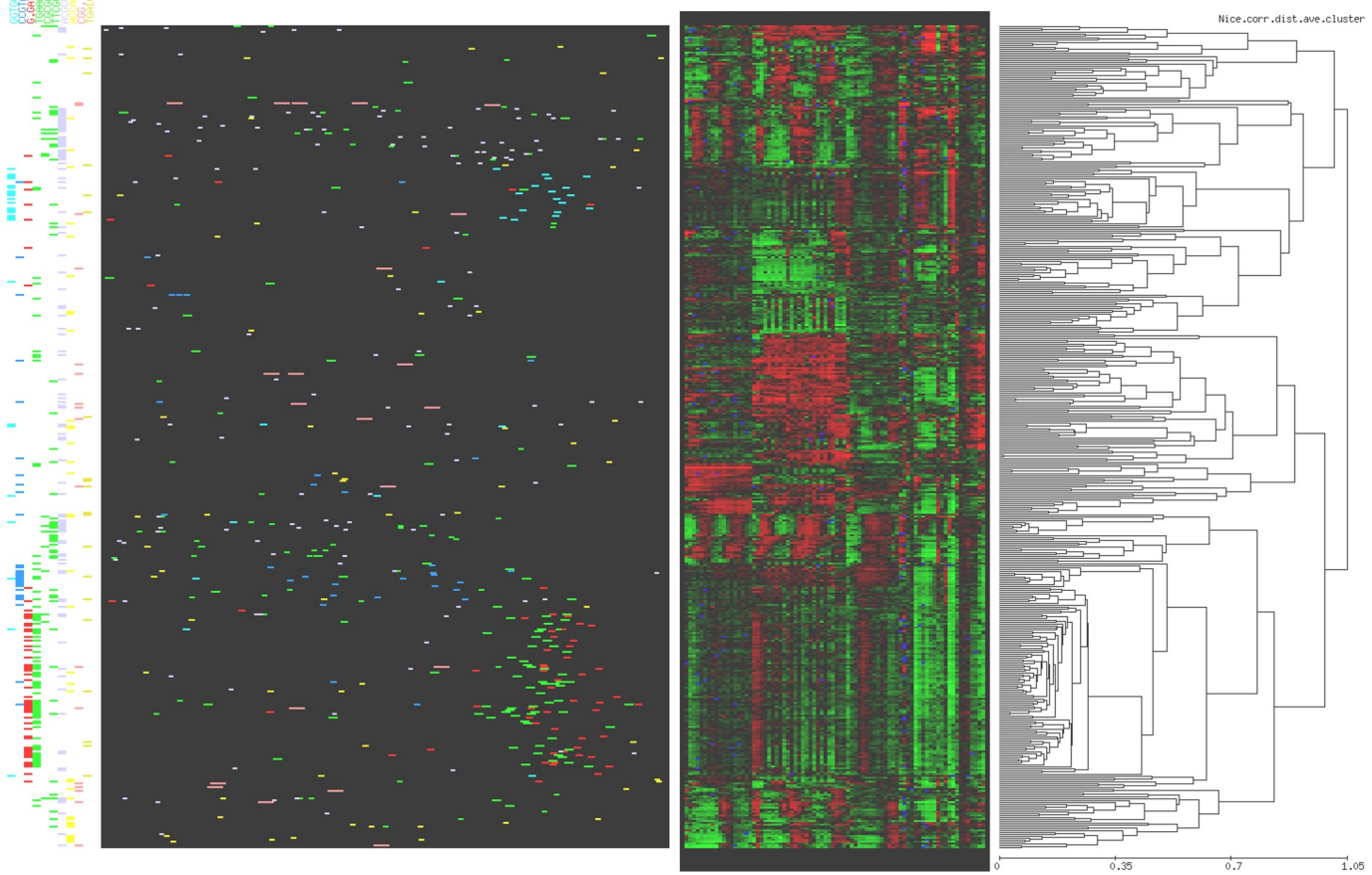
Converting from a mathematician to a biologist



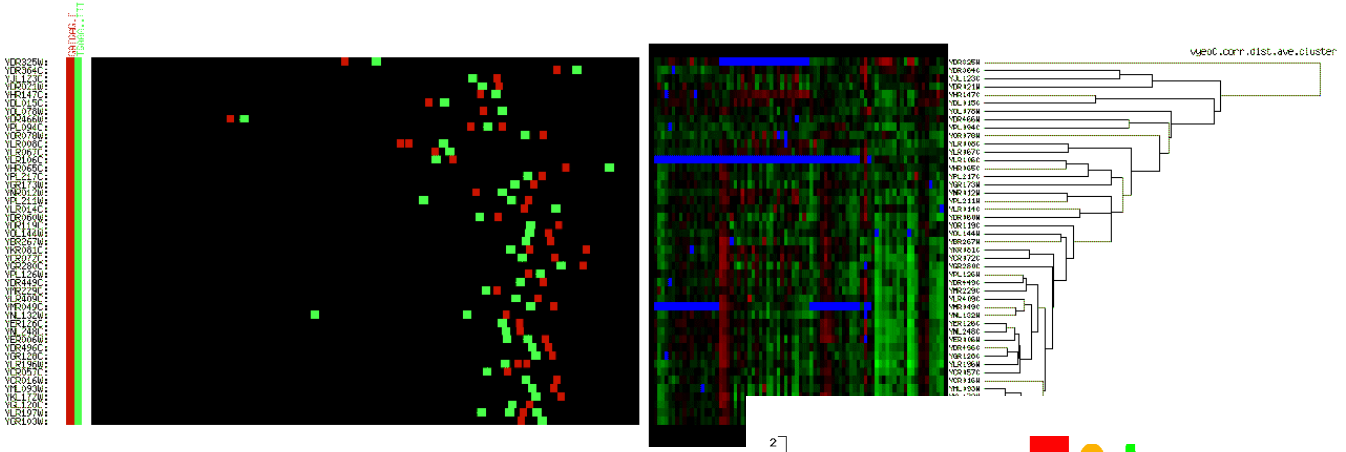
1997 – the first Eukaryotic genome ever sequenced



Microarrays – the beginning of functional genomics

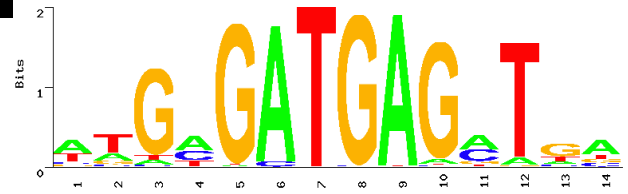

Nice.corr.dist.ave.cluster
0
0.35
0.7
1.05

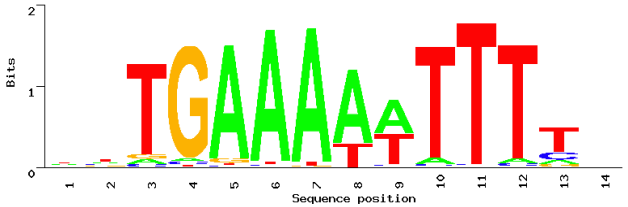
The hypothesis – genes that are similarly expressed should be similarly regulated



GATGAG.T

TGAAA..TTT





Brazma A, Jonassen I, Vilo J, Ukkonen E. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Research* (1998)

EMBL-EBI

If you have a hammer but you see a screw, put the hammer aside and look for a screwdriver.

(But do not forget where you have put your hammer, it may become handy one day)

Lessons learned

- Trying to *explain as much as possible with as little as possible* **is** relevant to biology
- If you have a hammer but you see a screw, put the hammer aside and look for a screwdriver. But do not forget where you have put your hammer, it may become handy one day
- It’s good to be in the field early!!!

ChatGPT, write one paragraph about **the future of bioinformatics** in the light of AI

The future of bioinformatics is highly intertwined with the use of artificial intelligence (AI). The ability of AI to learn and analyze vast amounts of biological data presents a significant opportunity to advance our understanding of complex biological systems. With the use of machine learning algorithms and other AI techniques, bioinformatics researchers can make sense of large datasets, identify novel patterns and correlations, and develop predictive models for a wide range of applications in healthcare and life sciences. Furthermore, AI can also enable the integration of diverse data sources, such as genomics, transcriptomics, proteomics, and metabolomics, allowing researchers to gain a more comprehensive understanding of biological systems. Overall, the integration of AI in bioinformatics is poised to revolutionize the field, opening up new avenues for discovery and enabling researchers to develop more personalized and effective treatments for various diseases.

continue 16, 32, 64, 128, 256, ...

512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, 131072, 262144, ...

continue 16, 1, 6, 32, 3, 2, 64, 6, 4, 128, 1, 2, 8, 256, 2, 5, 6, ...

This sequence doesn't seem to have an obvious pattern, but here are the next few terms:

512, 4, 8, 1024, 9, 10, 2048, 12, 16, 4096, 1, 18, 32, 8192, 20, 25, 18, 16384, 27, 4, 65536, 30, 32, 512, ...

Without more information, it's difficult to determine a clear rule or formula that generates this sequence. It's possible that the sequence is arbitrary or follows a complex pattern that is not immediately apparent.

To learn more what ChatGPT 4 can and cannot do, read my blog on LinkedIn