Sobre el uso de algoritmos evolutivos para encontrar leyes a partir de datos: Éxitos y límites

Emilio Hernández-García

*IFISC

The Automation of Science

Ross D. King,[1*] Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research, we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested treelike structure, 10 levels deep, th… description. This formalizati…

Computers are playing … in the scientific proc… control the execution … tributes to a vast expansion …

experiments can be executed, each individual experiment cannot be designed to test a hypothesis about a model. Robot scientists have the potential to overcome this fundamental limitation.

The complexity of biological systems necessitates the recording of experimental metadata in as much detail as possible. Acquiring these metadata has often proved problematic. With robot scientists, comprehensive metadata are produced as a natural by-product of the way they work. Because the experiments are conceived and executed automatically by computer, it is possible to completely capture and digitally curate all aspects of the scientific process (11, 12).

Distilling Free-Form Natural Laws from Experimental Data

Michael Schmidt[1] and Hod Lipson[2,3*]

For centuries, scientists have attempted to identify and document analytical laws that underlie physical phenomena in nature. Despite the prevalence of computing power, the process of finding natural laws and their corresponding equations has resisted automation. A key challenge to finding analytic relations automatically is defining algorithmically what makes a correlation in observed data important and insightful. We propose a principle for the identification of nontriviality. We demonstrated this approach by automatically searching motion-tracking data captured from various physical systems, ranging from simple harmonic oscillators to chaotic double-pendula. Without any prior knowledge about physics, kinematics, or geometry, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. The discovery rate accelerated as laws found for simpler systems were used to bootstrap explanations for more complex systems, gradually uncovering the "alphabet" used to describe those systems.

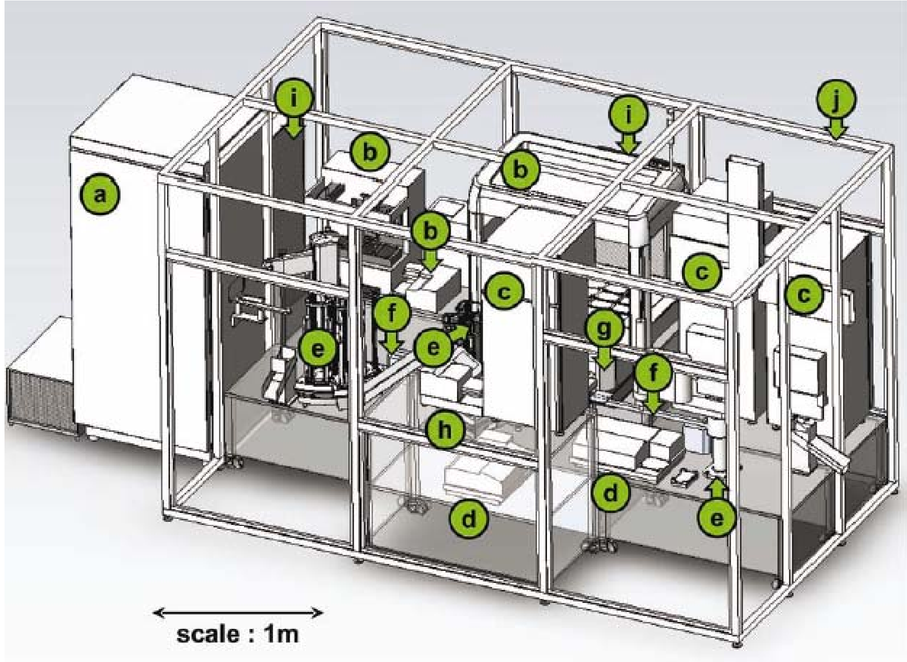
section S4 in the supporting onli… (SOM)]. Unlike traditional linear an… regression methods that fit parame… equation of a given form, symbolic… searches both the parameters and t… equations simultaneously (see SOM s… Initial expressions are formed by rand… bining mathematical building block… algebraic operators {+, −, ÷, ×}, … functions (for example, sine and co… stants, and state variables. New eq… formed by recombining previous equ… probabilistically varying their sube… The algorithm retains equations that… experimental data better than others… dons unpromising solutions. After equ… a desired level of accuracy, the algor… nates, returning a set of equations th… likely to correspond to the intrinsic r… underlying the observed system.

Science **324**, pp 81 and 85
3 April 2009

OUTLINE

- Motivation
- Automation of Science (Adam, the Robot Scientist)
- Finding natural laws from data
- Essentials of genetic algorithms
- Predicting time series from the Mediterranean sea
- Reflections, and outlook

http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/



Adam, the robot scientist
King et al. Science 324, 81 (2009)

What is ADAM?

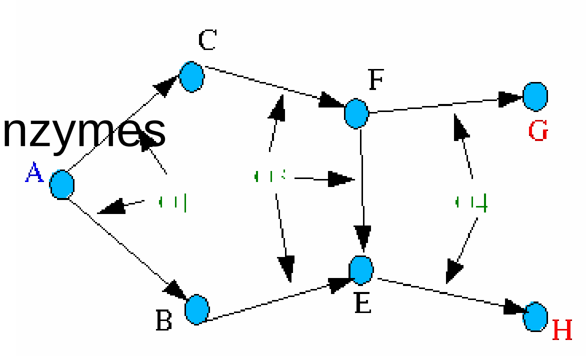
It is a fully automated laboratory to perform a kind of specialized task: **recording growth curves of different strains of yeast (**S. *cerevisiae***) mutants in different media** (thousands of strains, 6 metabolites)

It is controlled by a computer that has some ‘Artificial intelligence’: It is programmed to search its databases to deduce (abduce) hypothesis on **which yeast gene codifies some ‘orphan’ enzymes** (enzymes with unknown coding gene(s)). Then, it **plans experiments for checking the hypotheses**, and **performs** them.

Thus, in some sense, is like a scientist which runs a single experimental program (of thousands of experiments per day)

The knowledge inside Adam:

enzymes



metabolites

- Whelan & King 2008 graph model of yeast metabolism

- KEGG bioinformatic database (annotated genes and proteins from many organisms)

Simple heuristics for formulating hypotheses:

For the yeast orphan enzymes, select those affecting the end-point metabolites
Find their EC enzyme class
Look for genes in other organisms codifying for enzymes in this EC class
Search for genes in yeast homologous to those

Hypothesis: these homologous in yeast codify for the orphan enzyme

And experimental hypothesis testing:

Find available metabolites metabolically linked to the enzyme, and perform multiple experiments, measuring growth curves in wild-type and in the mutant lacking the candidate gene, in presence and absence of the metabolite, and reject or keep the hypothesis.

Discoveries by Adam:

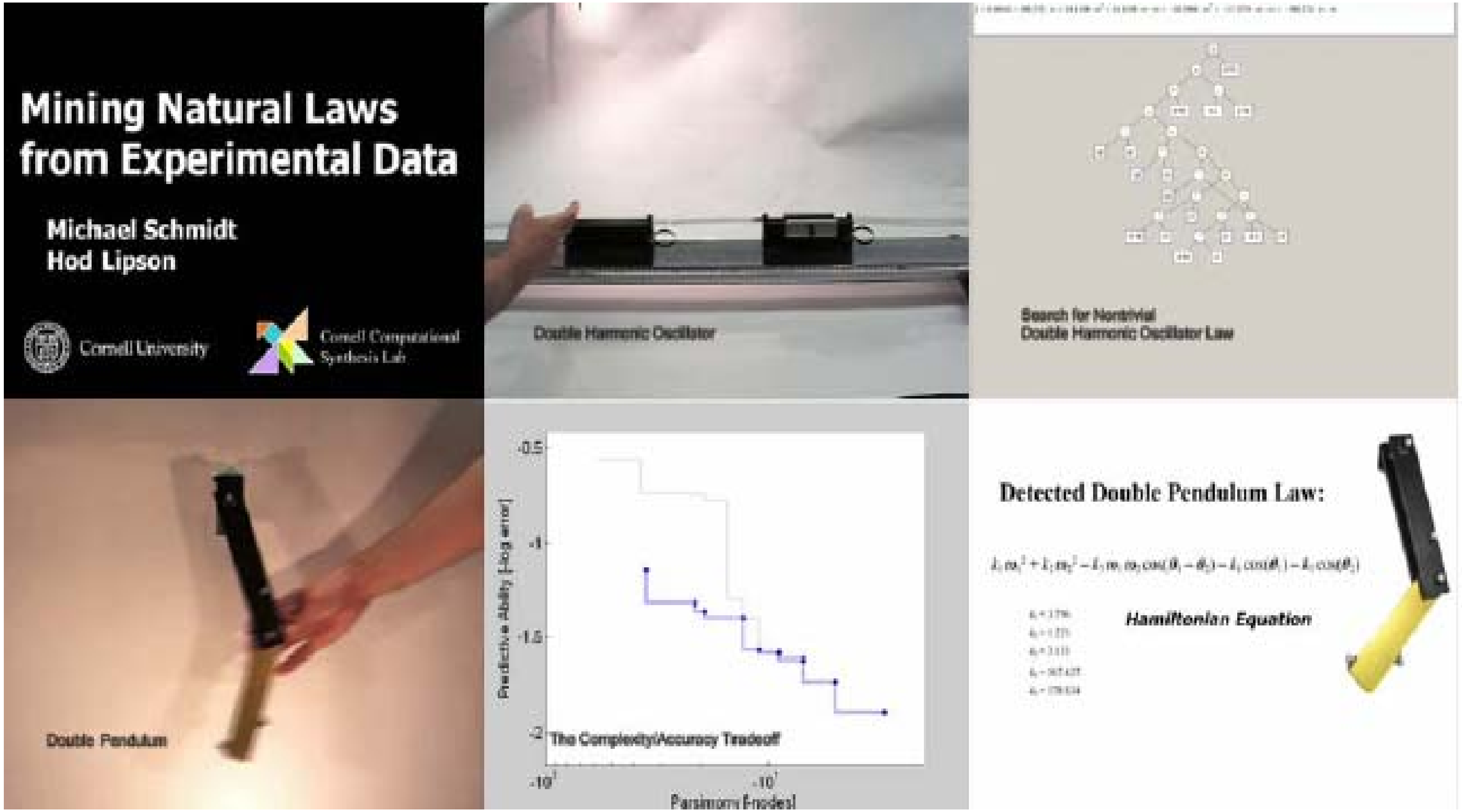
From 13 orphan enzymes, 20 hypothesis formulated, 12 confirmed. After revision, 6 were already identified in the literature (thus the function of the remaining 6 genes has been discovered by the Robot (and one possible error in the bioinformatic database pointed out)).

Example: genes YGL202W, YJL060W, and YER152C encode for the enzyme 2A2OA

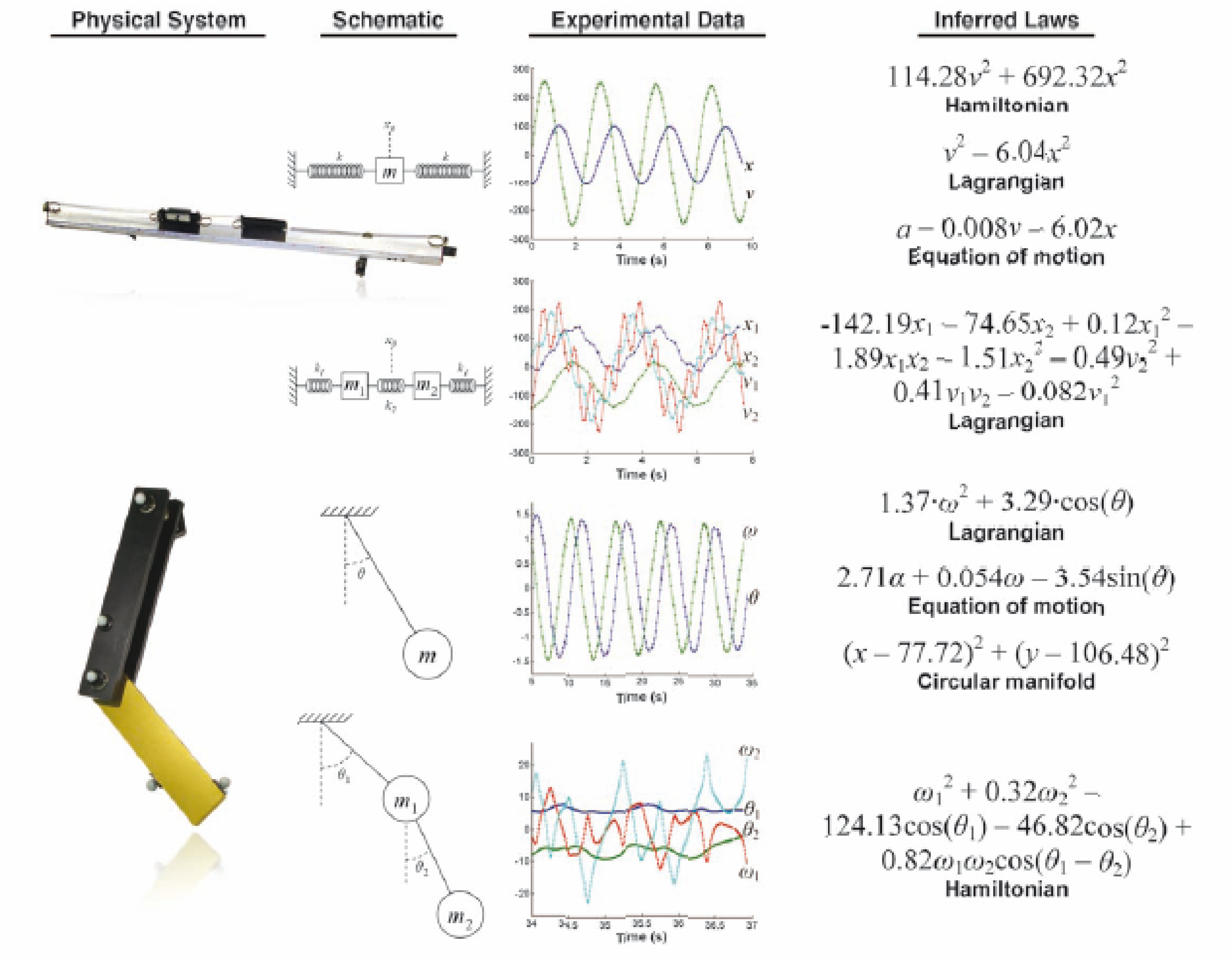
Is Adam really a scientist? well … he does some of the stuff some biological scientists do, at least earlier in their carrier …

It is essentially constrained by the initial knowledge in it: metabolism (fixed in its memory) and bioinformatics (only slightly improving by its discoveries). Is there any way to automate a more drastic improvement of the initial knowledge?

DestillingNaturalLaws.mpg



Schmidt and Lipson, Science 324, 81 (2009)

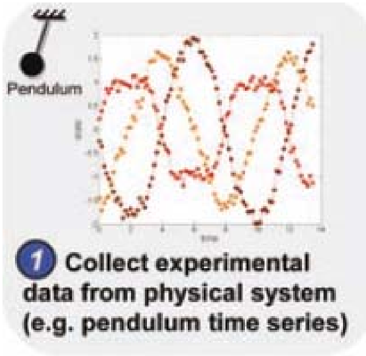


The computer algorithm, just from data, without any ‘database’ of knowledge on physics, has ‘discovered’ conservation laws that physicists identify as correct physical laws

The algorithm used by the computer to ‘learn’ laws from data is a
Genetic Algorithm (GA)

- **GAs:** Search and Optimization algorithms based on the mechanisms of biological evolution
- Developed by John Holland, University of Michigan (1970’s)
- Provide efficient, effective techniques for optimization and machine learning applications
- Widely-used today in business, scientific and engineering circles

GAs as applied to the "inferring laws from data" problem:



$$f = (x - 1.12) \cdot \cos(y)$$
$$f = 0.91 \cdot \exp(y/z)$$
$$f = 0.5 \cdot y^2 - 9.8 \cdot \cos(x)$$

2) Generate an initial population of many random formulae involving the variables measured (random physical laws)

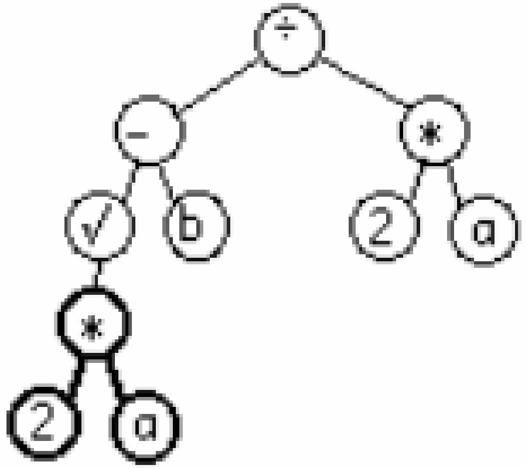
3) Apply each formula to the data to see if it gives a good fit or not. A fitness value is assigned to each formula, measuring how well the data agree with it.

4) Create a new generation of formulas:
   i) Copy the best existing formulas
   ii) Create new formulas by mutation of old ones
   iii) Create new formulas by crossover (sexual reproduction).

5) Repeat again and again. By this 'artificial selection' process, each generation contains formulas better than the previous. After several thousands of generations the formulae obtained are indeed very good and the best one can be though as an inferred "physical law"

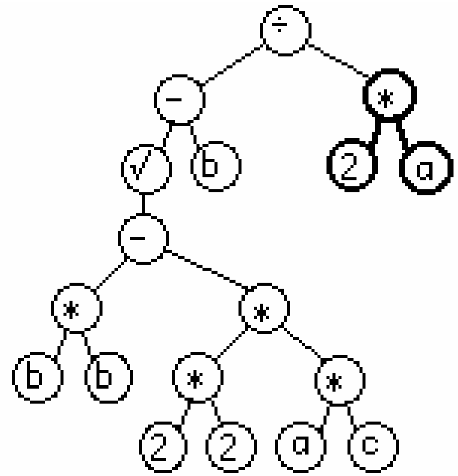
$$f = z + 9.8 \cdot \sin(x)$$
$$f = 0.5 \cdot y^2 - 9.8 \cdot \cos(x)$$

A convenient way to store a formula in a computer is as a network:



$$= \frac{\sqrt{2a} - b}{2a}$$

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} =$$

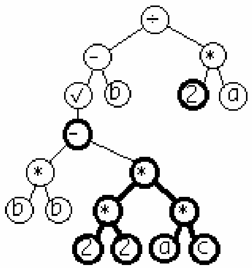


Thus, this implementation of GAs puts a population of networks to reproduce and compete until the “law” representing best the data is selected in this *struggle for being reproduced*

Mutation

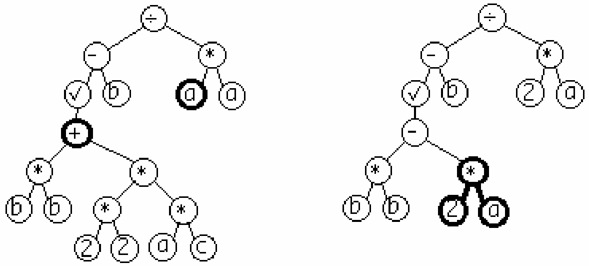
Original Individual

(÷ (- √ (- (* b b) (* (* 2 2) (* a c))) b) (* 2 a))



Mutated Individuals

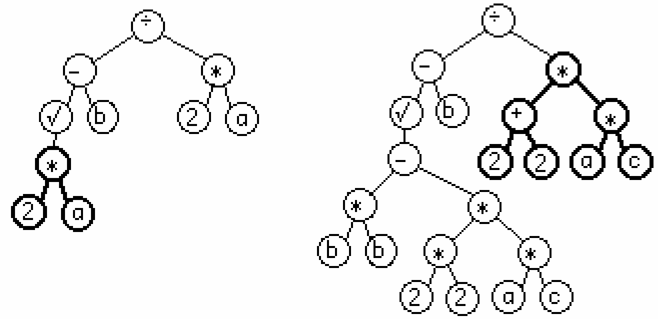
(÷ (- √ (+ (* b b) (* (* 2 2) (* a c))) b) (* a a))   (÷ (- √ (- (* b b) (* 2 a)) b) (* 2 a))



Crossover Operation with Different Parents

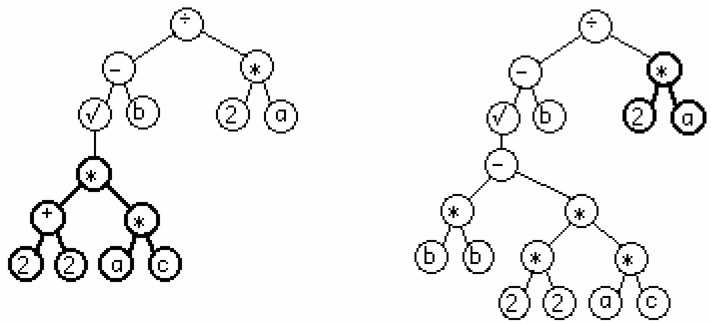
Parents

(÷ (- √ (* 2 a) b) (* 2 a))   (÷ (- √ (- (* b b) (* (* 2 2) (* a c))) b) (* (+ 2 2) (* a c)))



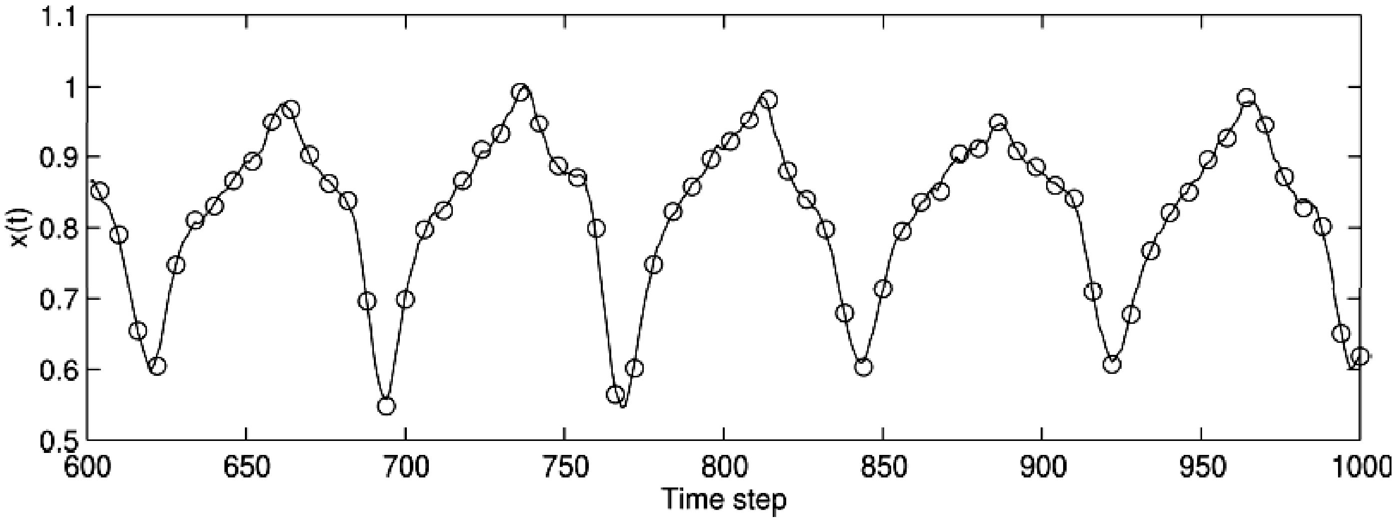
Children

(÷ (- √ (* (+ 2 2) (* a c)) b) (* 2 a))   (÷ (- √ (- (* b b) (* (* 2 2) (* a c))) b) (* 2 a))



$$\frac{\sqrt{b*b-2*2*a*c}-b}{2*a}=\frac{-b+\sqrt{b^2-4ac}}{2a}$$

Prediction of time series by GAs:



=35 s

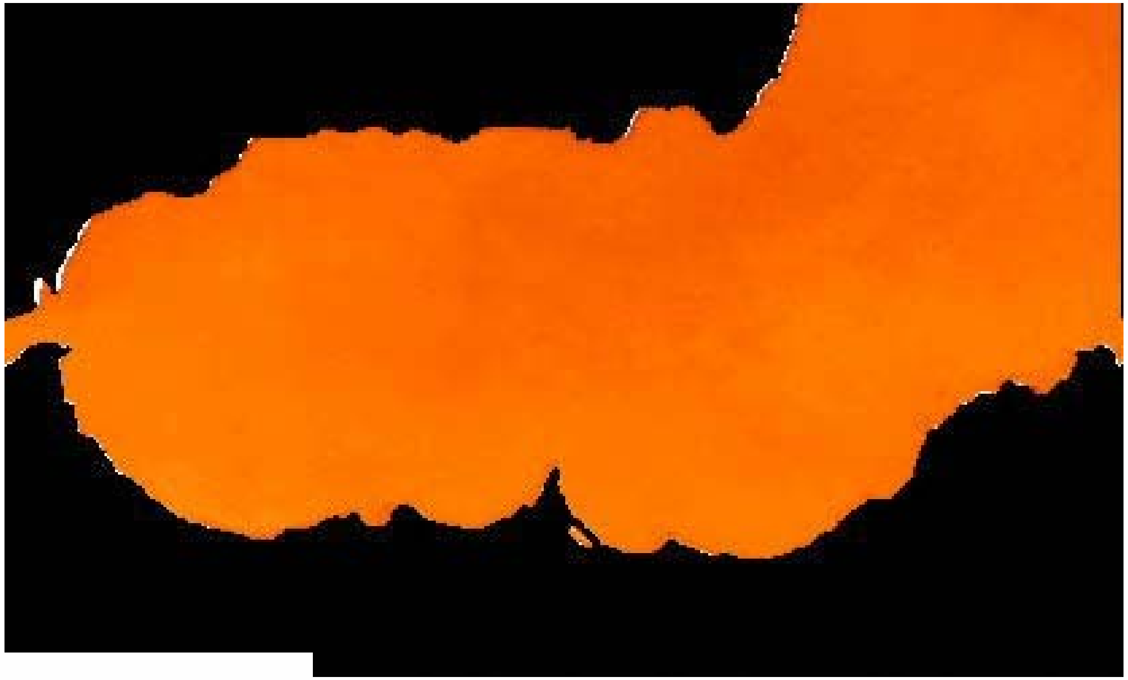
solid: Acceleration of the hand of a patient during Parkinson tremor
circles: one-step ahead prediction from the formula obtained by GA:

$$x_0(t) = (x_0(t-1) + ((x_0(t-3) * (x_0(t-1) - x_0(t-3))) * ((x_0(t-3)/(x_0(t-2)/((2.20) * x_0(t-1)))) * ((1.12) - x_0(t-2)))))$$

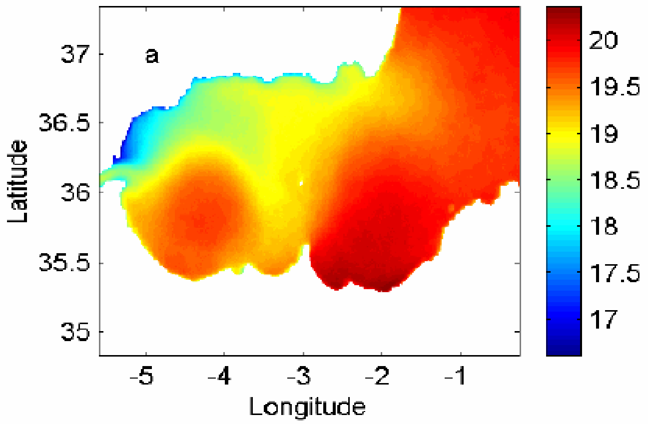
(Alvarez et al. Comp. Phys. Comm., 2001)

Prediction of the Sea Surface Temperature dynamics at the Alboran sea

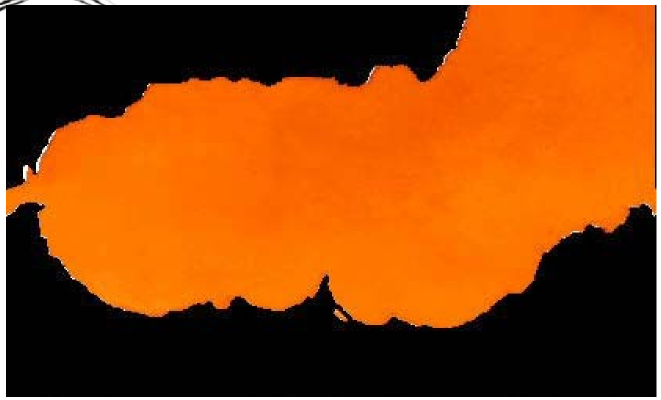
(Alvarez, Lopez, Riera, Hernandez-Garcia, Tintore, Geophys. Res. Lett. 27, 2709 (2000))



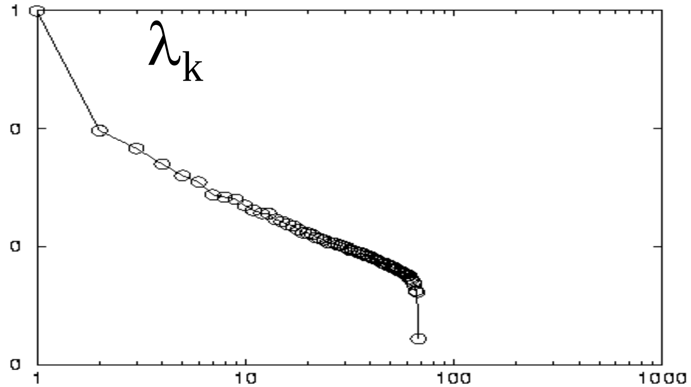
Decomposition into Empirical Orthogonal Eigenfunctions (or Principal Components)



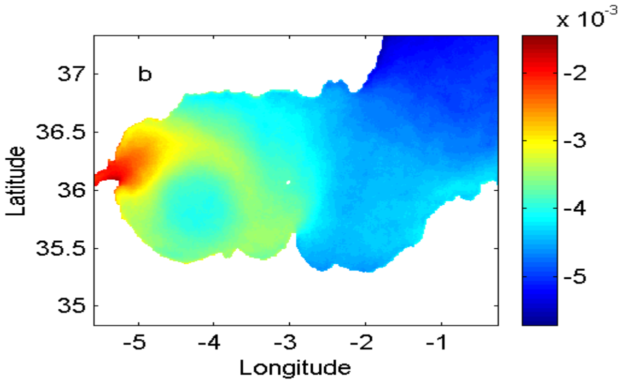
$415 \times 250 \approx 10^5$ temperature time series from satellite sensors (6 years of montly data)

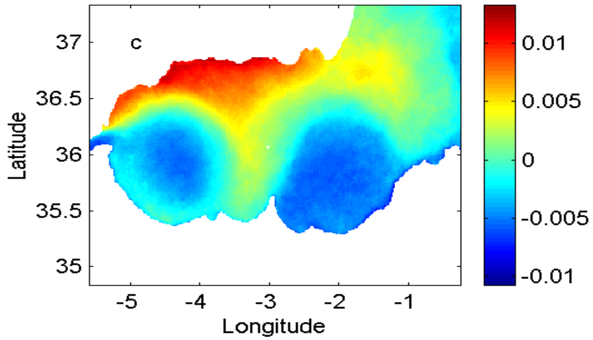


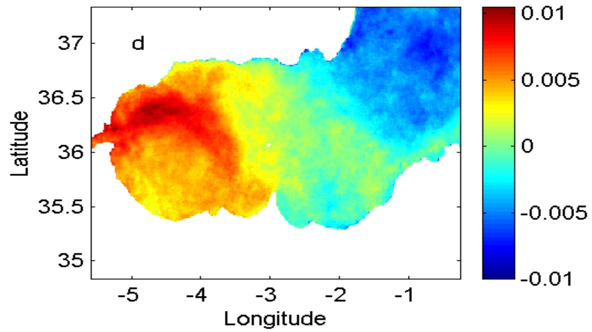
Alboran Sea SST



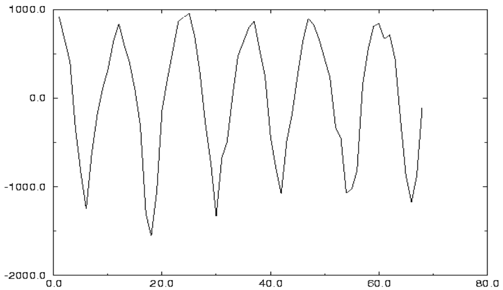
EOFs



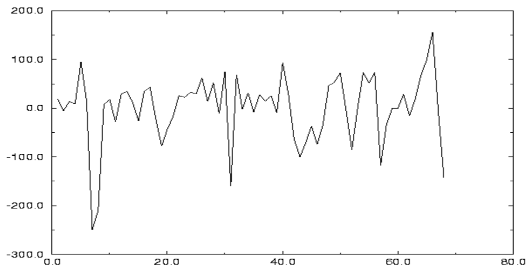




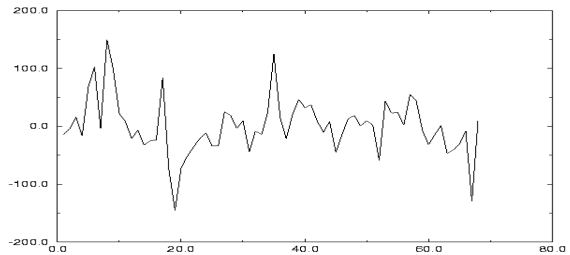
$a_t(1)$



$a_t(2)$

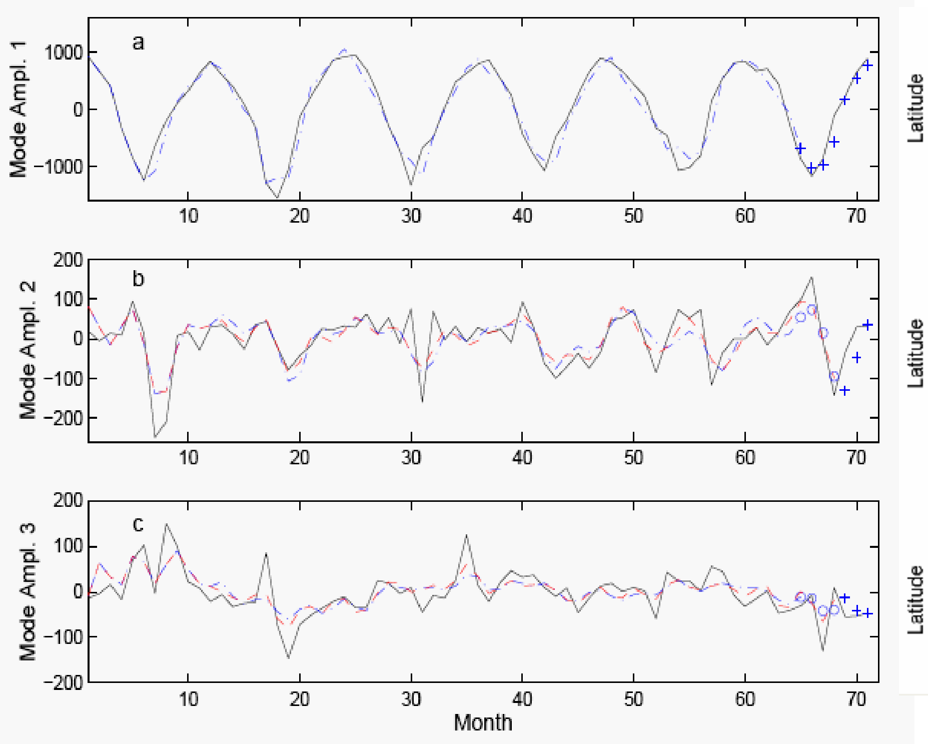


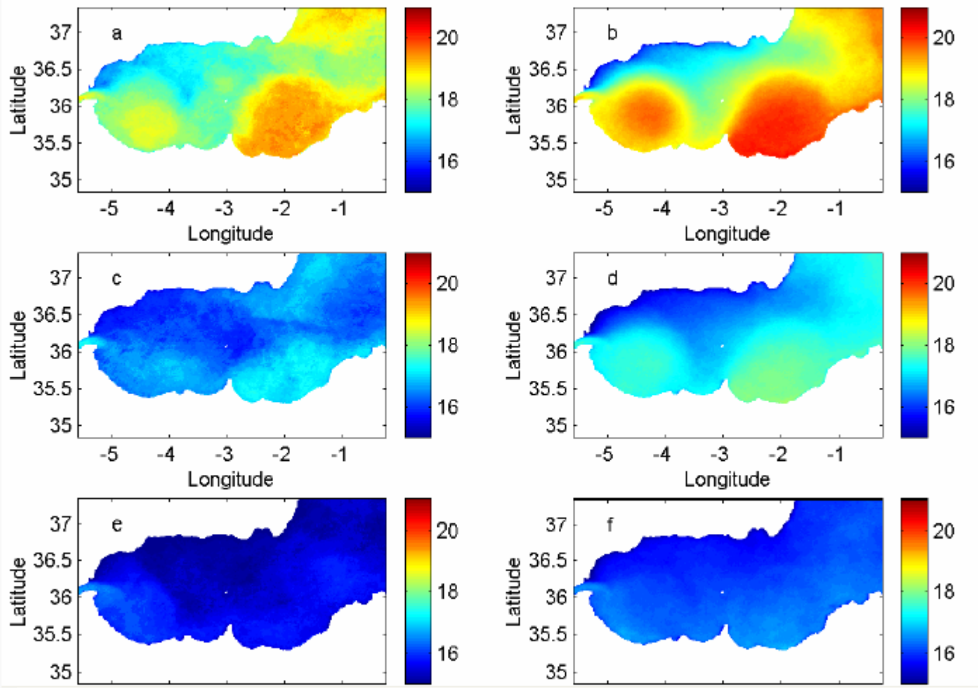
$a_t(3)$



Observed

Prediction 1 month in advance



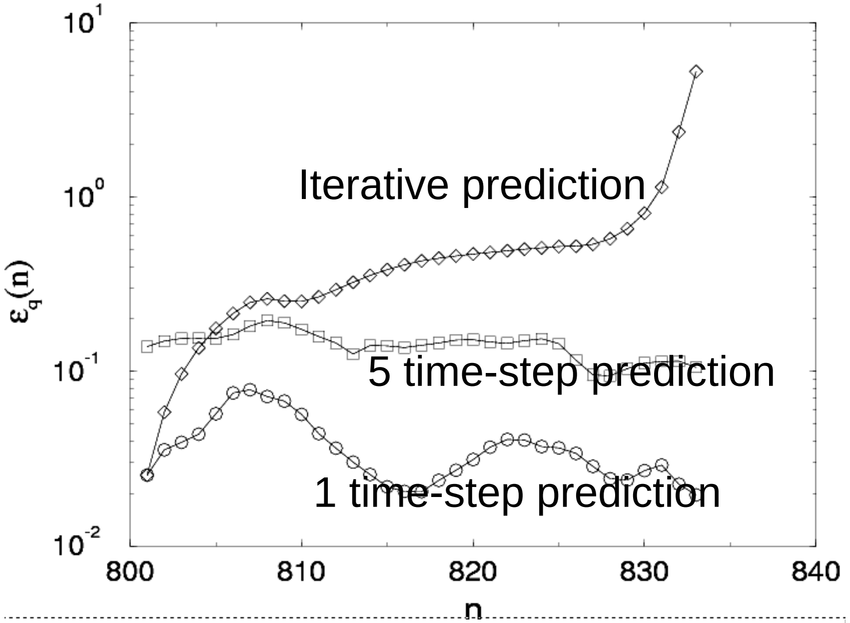


November 1998, December 1998, January 1999

OUTCOME, LIMITATIONS …

- GAs find nearly perfect predictor formulae for processes involving periodic or close to periodic oscillations
- Agreement is not so perfect for chaotic or turbulent motions. Lots of data and computer time to find reasonable "laws".
- Short time prediction relatively good, but longer time …

Prediction of virtual data from a model of turbulence



Note: it is very interesting to understand all of this in terms of predictability, chaos, noise, and the like, but not the subject of this talk

So, is this the kind of methodology able to provide “artificial intelligence” to computers or robots so that they can discover new physical, biological, … laws from data ?

Look at the expressions for the “laws” of temperature at the Alboran sea:

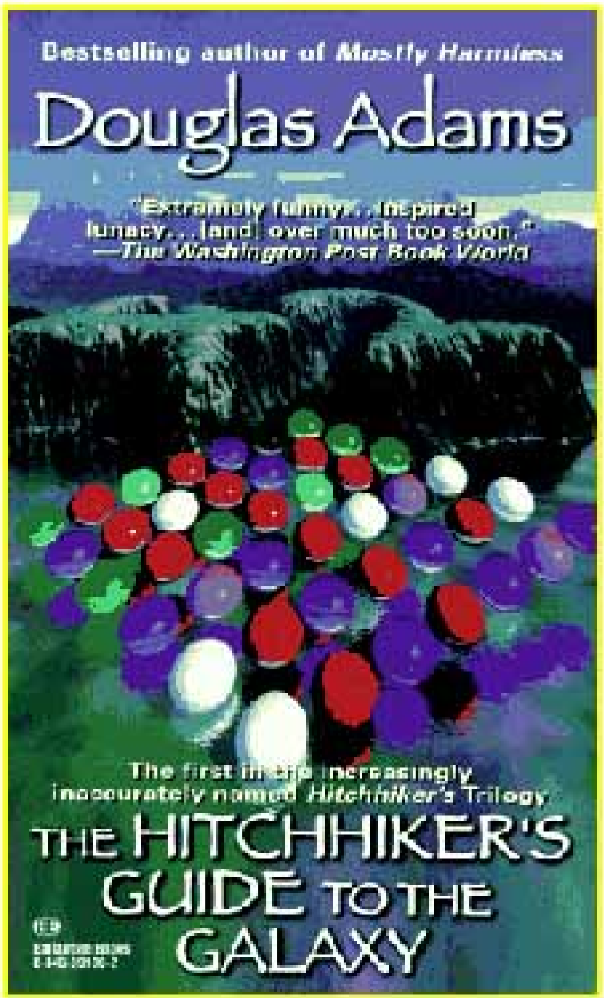
$$A_1(t) = 0.33\,\{2\;A_1(1) - [A_1(3) + A_1(6) + (A_1(1)\,\left[A_1(2)^{-1}\,(9.3 - A_1(1)) - 3.78\right]^{-1}\Big)\Big]\Big\}. \qquad \text{(A1)}$$

$$A_2(t) = A_2(1) - A_2(2) - 0.134\,\{A_2(4) - (A_2(5) - A_2(12) - 3.45\,[A_2(5) + A_2(8)])\}. \qquad \text{(A2)}$$

$$A_3(t) = 0.4A_3(12) - 0.4 - 0.59\,[2.5 - A_3(3) + A_3(9) - A_3(1)]. \qquad \text{(A3)}$$

Can this be called “physical law”?

In the Schmidt and Lipson study structures where recognized because the laws were discovered by human research centuries ago



The answer to the Ultimate Question of Life, the Universe, and Everything …

This way of discovering “laws” from data is a practical way of implementing the inductivist vision of science defended by Bacon (1561-1626) o Mill (1806-1873). Today, most scientists agree that the scientific method is not just compiling data in compressed forms (what Rutherford called *stamp collecting*):

"Lo que hizo posible el análisis de la multiplicación bacteriófaga, y una compresión de sus diferentes etapas, fue por encima de todo el juego de hipótesis y experimentos, construcciones de la imaginación e inferencias que se podían extraer de ellas. Comenzando por una cierta concepción del sistema, se ideaba un experimento para poner a prueba uno u otro aspecto de esta concepción. En función de los resultados, se modificaba la concepción para proyecta otro experimento. Y así sucesivamente y sucesivamente. Así es cómo funcionaba la investigación en biología. En contra de lo que yo antaño pensaba, el progreso científico no consistía simplemente en observar, en acumular hechos experimentales y extraer una teoría a partir de ellos. Comenzaba con la invención de un mundo posible, o un fragmento de él, que luego se comparaba con el mundo real a través de la experimentación. Y era este diálogo constante entre la imaginación y el experimento lo que permitía que uno se formase una concepción cada vez más fina de lo que se llama realidad".

François Jacob,.1964.*The Statue Within*.NewYork:Harcourt,Brace,Word

- Genetic algorithms (and in fact other techniques from the Artificial Intelligence area such as neural networks, etc.) provide extremely powerful and automated methods for machines to accumulate observed knowledge into very compressed form, and this is very useful (and used) in control of industrial processes, optimization, design, prediction of risk situations, and for sure in biological applications …

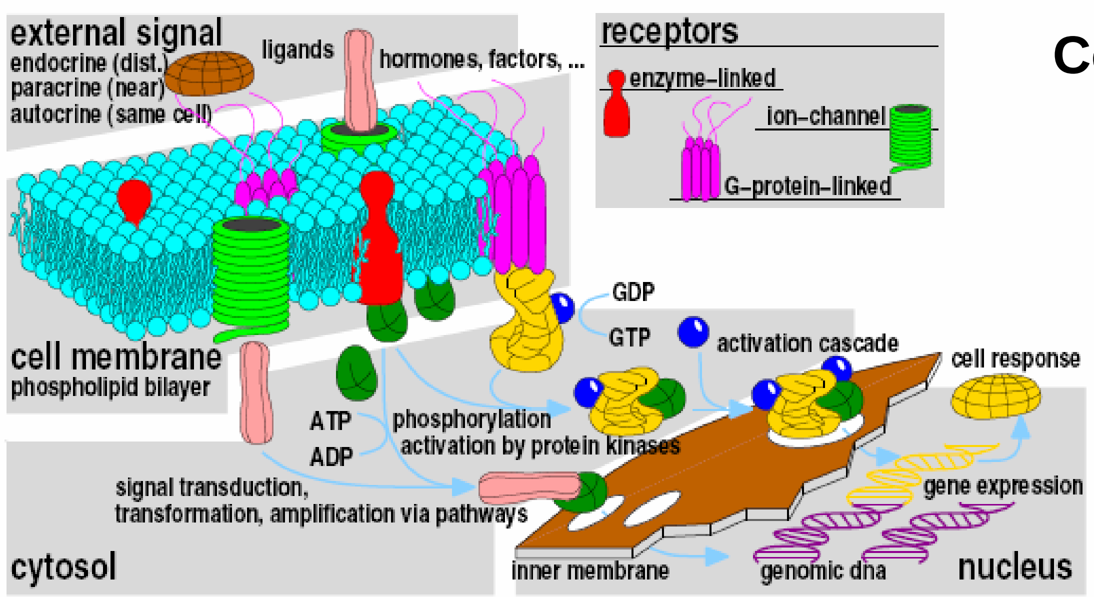
- The fact that observations can be compressed indeed indicates that there is some “natural law” out there. But calling “natural law” to the direct output of GAs seems excessive, and only justified in very simple cases such as the ones studied by Schmidt and Lipson.

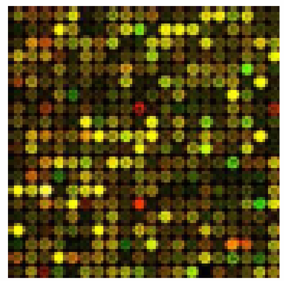
- Instead, the “empirical laws” found by automated methods may provide clues on the true “natural laws” (for example focusing on the “motifs” which repeat in the formulae obtained).

- The case of Adam is more related to a true scientific method, in which the robot explores its internal “knowledge” and models, formulates hypothesis and tests them. But feedback from the experiment onto its internal knowledge –learning- seems still much more restricted than in human scientists.

IFISC

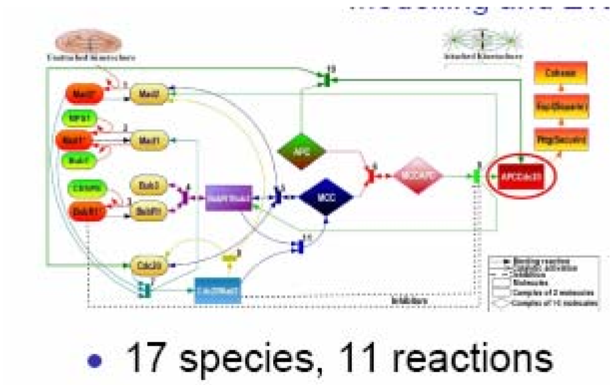
Cell signaling





gene expression data visualised by microarray (TU Dresden, BIOTEC)

Human spindle assembly checkpoint



Evolving candidate signaling networks until agreeing with data seems a natural application of GAs to biological research. It would open the door to dentifying the basic modules and pathways

Lenzer, Hinze, Ibrahim, Dittrich, www.minet.uni-jena.de/csb
Evolutionary Network Reconstruction Tools