Metabolomics, machine learning and modelling: towards an understanding of the language of cells

Douglas Kell

School of Chemistry, University of Manchester, MANCHESTER M60 1QD, U.K.

dbk@man.ac.uk

http://dbk.ch.umist.ac.uk/

http://www.mib.ac.uk

MANCHESTER 1824 The University of Manchester

Aber Genomic Computing

Understanding complex systems and how their subsystems communicate

Douglas Kell

School of Chemistry, University of Manchester, MANCHESTER M60 1QD, U.K.

dbk@man.ac.uk

http://dbk.ch.umist.ac.uk/

http://www.mib.ac.uk

MANCHESTER 1824 The University of Manchester

Aber Genomic Computing

"Progress in science depends on new techniques, new discoveries, and new ideas, probably in that order"

Sydney Brenner, Nature, June 5, 1980

"But one thing is certain: to understand the whole you must study the whole"

Henrik Kacser, 1986

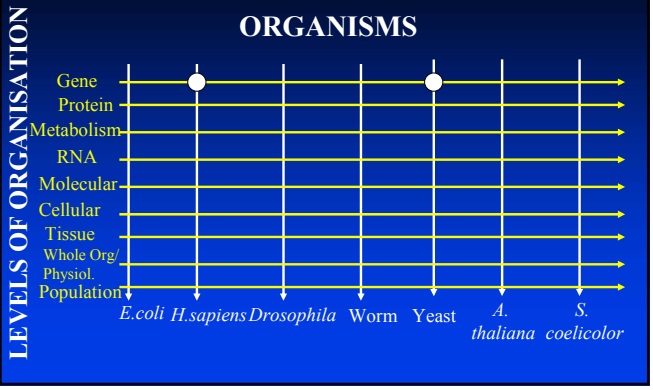
SCOPE OF THE TALK

New ideas, new techniques and new developments

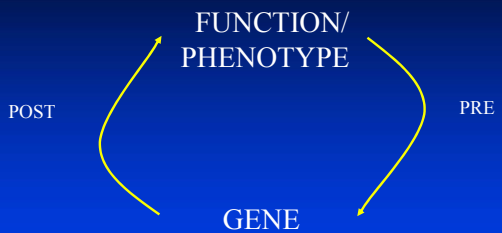
- Philosophical elements of systems biology
- Genetic programming analysis of the metabolome
- Metabolic footprinting – a novel strategy for functional genomics and mode of action studies
- Tuning mass spectrometers via genetic search
- Metabolomics for disease diagnostics
- Systems Biology of the NF-κB signalling pathway
- Cell Signalling as signal processing

- MIB and Conclusions

A Potted History of Biology



Pre- and post-genomics



Holism/reductionism

WHOLE (ORGANISM)

SYNTHESIS/ HOLISM

REDUCTIONISM

PARTS (MOLECULES)

Holism/reductionism

WHOLE (ORGANISM)

QUANTITATIVE

QUALITATIVE

PARTS (MOLECULES)

Lazebnik – how a biologist learned to fix a radio

Cancer Cell 2, 179-182 (2002)

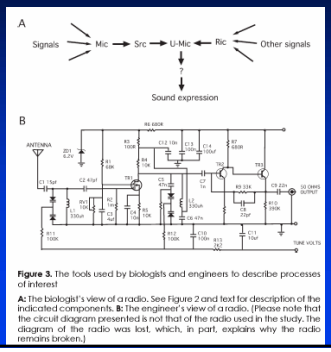


Figure 3. The tools used by biologists and engineers to describe processes of interest

A: The biologist's view of a radio. See Figure 2 and text for description of the indicated components. **B:** The engineer's view of a radio. (Please note that the circuit diagram presented is not that of the radio used in the study. The diagram of the radio was lost, which, in part, explains why the radio remains broken.)

Models and Reality

THE BIOLOGY

RUN THE MODEL

PRODUCE/ REFINE THE MODEL

THE *IN SILICO* MODEL

Modelling

PARAMETERS

INFERENCE/ SYSTEM IDENTIFICATION

PRODUCE/ REFINE e.g. O.D.E. MODEL

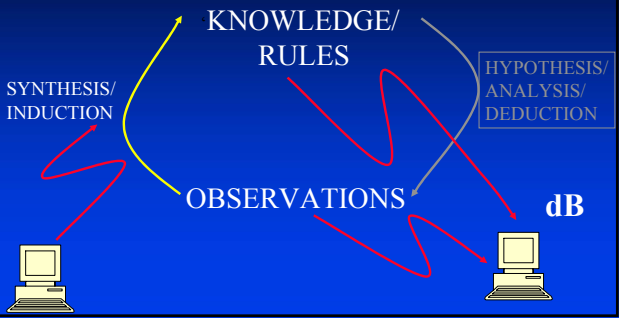
VARIABLES

Theory and Experiment

THEORY

EXPERIMENT

The cycle of knowledge



Timeline

- Pre-genomics → Post-genomics/ functional genomics → Systems Biology
- Organismal → Cellular → Molecular → Systems
- Forerunners, e.g. in Metabolic Control Analysis, Metabolic Engineering, Systems Theory, Synergetics

An important stress on the role of technology development in advancing science

• Almost Anything we can do now is the result of advances in technology, which have no 'hypothesis' (beyond the view that such an ability would be valuable)

•Lest you doubt it, a few examples from the modern era

e.g…..(most with Nobel prizes)

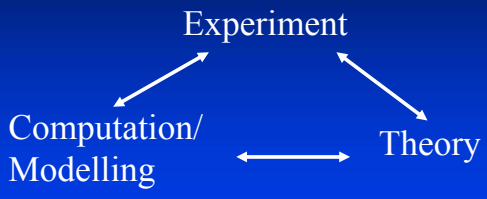
- Restriction enzymes
- Sequencing DNA and proteins
- PCR
- Computation and Internet for bioinformatics and even literature access
- Mass spectrometry for proteomics
- Voltage clamp & patch clamp for neurophysiology

Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era

Douglas B. Kell[1]* and Stephen G. Oliver[2]

BioEssays 26:99–105, © 2003 Wiley Periodicals, Inc. BioEssays 26.1 99

One view of systems biology



‘Bottom-up’ Systems Biology pipeline (dry)

1. Qualitative (‘structural’) model – who talks to whom as substrate, product or effector →
2. Quantitative model including ‘real’ or approximate equations describing individual steps →
3. Parametrisation of those equations →
4. Run the model and assess its most important parameters
5. Iteratively , with wet data, GOTO 1….

Systems biology experiments (including the wet side) ….

- Set up a well-defined system
- Effect systematic perturbations (genetic, environmental, chemical)
- Measure a time series of as many concentrations of variables, especially RNAs, proteins, metabolites (the ’omes) as possible
- Model the system and compare the experimental time series to those generated by the model
- Repeat iteratively

Post-genomics

- A chief result of the systematic genome sequence programs was the discovery of huge numbers of genes whose existence - let alone function - had previously gone unrecorded
- Post-genomics methods (aka ‘functional genomics’) are designed to find out the function of such genes, often by global ’omics methods in ‘known’ and ‘unknown’ genetically defined strains

A specific first aspect relates to determining which players are involved in any particular cellular process (starting to make the structural model)

’omics methods for high-information-content analysis

- Genome
- Transcriptome
- Proteome
- Metabolome ←
- Physiome
- etc.

Why metabolomics?

(i) ‘downstream’ - changes in the metabolome (metabolite concentrations, not fluxes – see MCA tutorial on our website for ‘why’) are amplified relative to changes in the transcriptome and the proteome, and are numerically more tractable,

(ii) no need for a whole genome sequence or a large EST databases to be available for each species,

(iii) metabolic profiling is much cheaper and very much more high-throughput than are proteomics and transcriptomics, making it feasible to examine large numbers of samples from organisms that have been ‘grown’ under a wide range of conditions,

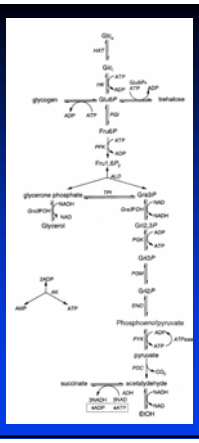
(iv) technology is generic as a given metabolite - unlike a transcript or protein - is the same in every organism that contains it (of course this is not true for secondary metabolites…),

(v) such methods have already been demonstrated.

The ideal analytical method

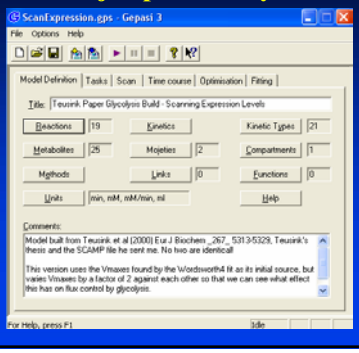
- Specific or highly selective
- Precise Accurate Reproducible
- Rapid Sensitive Non-destructive
- Low cost Reagentless / Probes biologically inert
- Robust equipment
- Easy to set up and calibrate
- Capable of axenic operation
- Signals linear with determinand concentration
- Global in scope (for high-information-content ’omics methods); → no prejudgement of ‘the answer’
- User-intelligible output ←

A glycolytic ‘pathway’, where we nominally know the metabolites and the enzymes involved



And if we know their parameters……

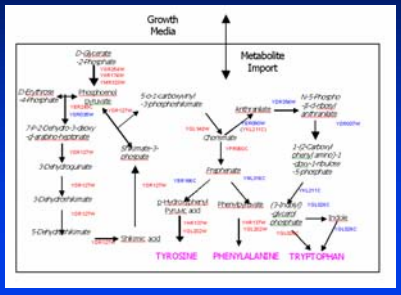
Simulating metabolism in the ‘forward’ direction – a major part of ‘Systems Biology’

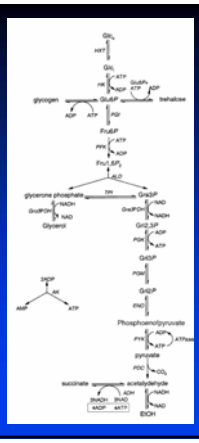


ODE-based modelling

- But for this (bottom-up) approach we need to know the parameters such as binding and rate constants, and these are not measured using ’omics strategies
- Omics measurements are measurements of variables, and variables cannot control other variables (nor parameters)

Omics lead to a ‘network-finding’, inverse problem





These methods do not scale well, the number of available samples is often small, and the problems are ill-posed

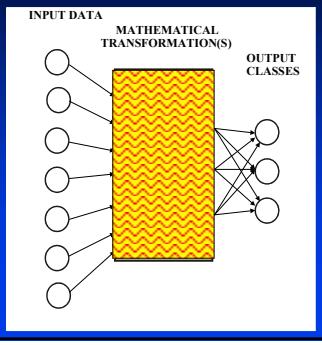
- Better is to divide the problem up into smaller problems to see the main players

Basic structure of propositional systems

Variables going across in different columns

Objects going down in different rows	X-var 1	X-var 2	X-var 3	Y-var 1	Y-var 2
Sample 1					
Sample 2…					

The machine learning paradigm



The combinatorial optimization problem

- Making a predictive model using n x-variables to predict just 1 y-variable gives 2^n models in which each one is used or not, before we even parametrise it, which is OK…but….
- …if n = 100, $2^n = 2^{100} \sim 10^{30}$; the lifetime of the Universe in seconds ~ 10^{17}…..
- If each variable can take just 10 values this is 10^n, etc…
- Machine learning methods are designed to search these huge spaces effectively
- …but in particular…→

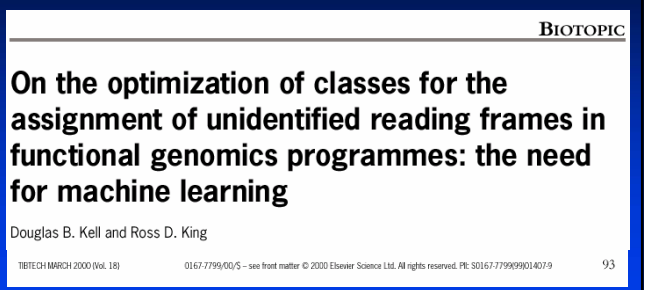
The combinatorial optimization problem

- Making a predictive model using n x-variables to predict just 1 y-variable gives 2^n models in which each one is used or not, before we even parametrise it, which is OK…but….
- …if n = 100, $2^n = 2^{100} \sim 10^{30}$; the lifetime of the Universe in seconds ~ 10^{17}…..
- If each variable can take just 10 values this is 10^n, etc…
- Machine learning methods are designed to search these huge spaces effectively
- ….the number of combinations if we only allow it to use 1,2,3,4 or 5 variables is just 100, 4950, 1.6 x 10^5, 3.9 x 10^6 and 7.5 x 10^7. These are much more tractable numbers, and are also likely to provide comprehensible explanations (and see later)

Some chemometric and related methods

Unsupervised *Just work on x-data*	*Supervised* *use y-data too*
• Principal Components analysis	• Back-prop neural networks
• Clustering methods	• Partial least squares regression
• Kohonen neural networks	
• Canonical variates analysis	• Discriminant Function Analysis
• Genetic Algorithms	• Inductive Logic Programming
• **Genetic programming**	
• Classification & Regression trees	

The functional genomics agenda (*inter alia*) is a supervised learning problem in which we use data from genes of known function to 'calibrate' those of unknown function



BIOTOPIC

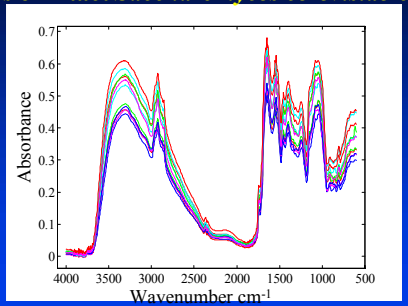
On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning

Douglas B. Kell and Ross D. King

TIBTECH MARCH 2000 (Vol. 18) 0167-7799/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0167-7799(99)01407-9 93

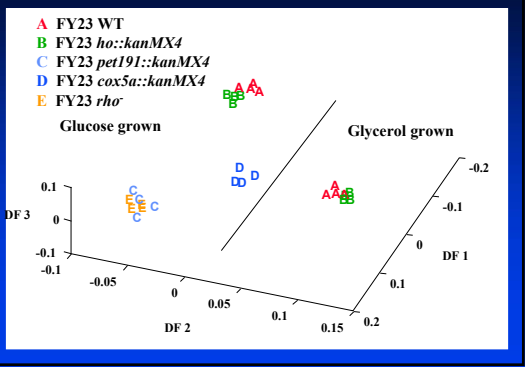
Tibtech 18, 93-98 (2000)

Diffuse reflectance-absorbance infra-red spectra of glucose-grown wild-type and single-gene knockout strains of intact *Saccharomyces cerevisiae* cells



MK Winson, DB Kell, F Baganz, SG Oliver
Tibtech 1998

Functional Genomics via Discriminant Function Analysis of FT-IR spectra of intact *Saccharomyces cerevisiae* cells



MK Winson, DB Kell, F Baganz, SG Oliver Tibtech 1998

Nature Biotechnol. (2001) 19, 45-50

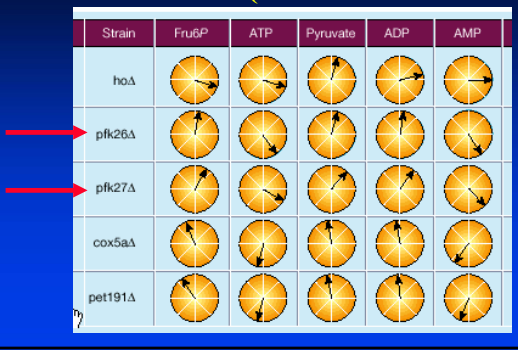
RESEARCH ARTICLES

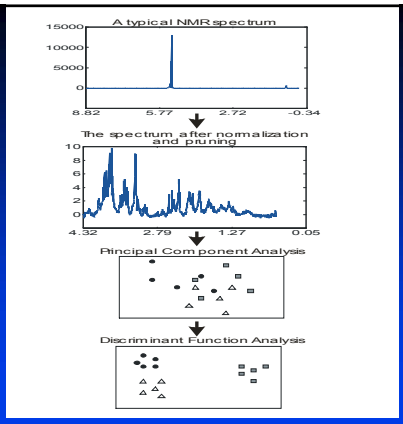
A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations

Léonie M. Raamsdonk[1], Bas Teusink[1,2], David Broadhurst[3], Nianshu Zhang[4], Andrew Hayes[4], Michael J. Walsh[1,5], Jan A. Berden[1], Kevin M. Brindle[6], Douglas B. Kell[3], Jem J. Rowland[7], Hans V. Westerhoff[1,8], Karel van Dam[1], and Stephen G. Oliver[4*]

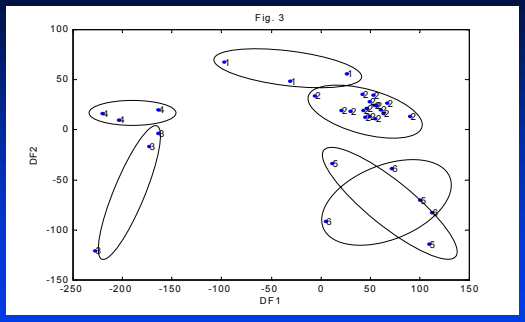
[1]Swammerdam Institute for Life Sciences, BioCentrum Amsterdam, University of Amsterdam, Plantage Muidergracht 12, NL-1018 TV Amsterdam, The Netherlands. [2]Current address: TNO-Prevention and Health, Zernikedreef 9, NL-2333 CK Leiden, The Netherlands. [3]Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, Aberystwyth SY23 3DD, UK. [4]School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester M13 9PT, UK. [5]Current address: Heineken Technical Services B.V., P.O. Box 510, NL-2380 BB Zoeterwoude, The Netherlands. [6]Department of Biochemistry, University of Cambridge, Old Addenbrooke's Site, 80 Tennis Court Road, Cambridge CB2 1GA, UK. [7]Department of Computer Science, University of Wales, Aberystwyth, Aberystwyth SY23 3DB, UK. [8]Department of Molecular Cell Physiology, BioCentrum Amsterdam, Free University, De Boelelaan 1087, NL-1081 HV, Amsterdam, The Netherlands. *Corresponding author (steve.oliver@man.ac.uk).

Co-response analysis of metabolites relative to G6P (arc cot of ratio of ln of changes)





Clustering of yeast functional genomics mutants – but - no 'rule' and no explanatory power

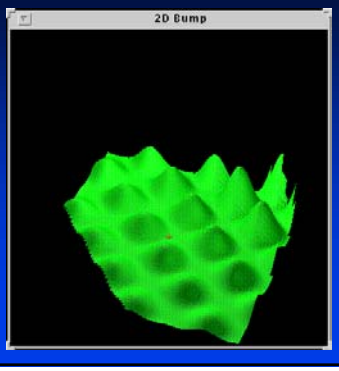


1) cox5a; (2) H0; (3) ρ0; (4) pet191; (5) pfk26; (6) pfk27.

Evolutionary computing (subsets include Genetic Algorithms, Evolutionary Strategies, Evolutionary Programs, Genetic Programming)

1. A population of individuals, each encoding a particular solution to a problem
2. A 'fitness function', by which we can evaluate how good that solution is (together these represent the 'landscape'…)

A scientific / combinatorial landscape



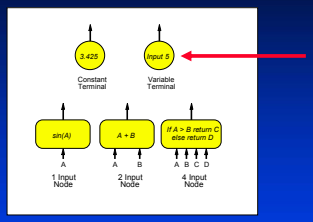
Red*: Simplex hill-climber
Yellow – Hooke & Jeeves
Purple – genetic algorithm
White - GA members
Redline – simulated annealing

EVOLUTIONARY ALGORITHMS

1. A population of individuals, each encoding a particular solution to a problem
2. A 'fitness function', by which we can evaluate how good that solution is (together these represent the 'landscape'…)
3. A selection strategy for determining who contributes to the next generation
4. Introduction of genetic diversity by e.g. mutation and recombination
5. A stopping criterion.

ALGORITHM CYCLES THROUGH STEPS 1 TO 4 UNTIL 5 IS SATISFIED

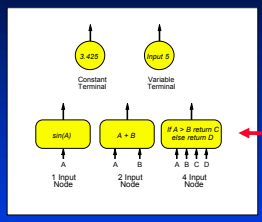
GP/GC BUILDING BLOCKS



A GP (e.g. Koza 1992) has two types of 'gene'

Terminals: numerical constants or input (x-) variables.
Nodes: mathematical operators or program functions.

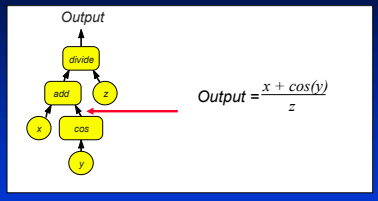
GP/GC BUILDING BLOCKS



A GP (e.g. Koza 1992) has two types of 'gene'

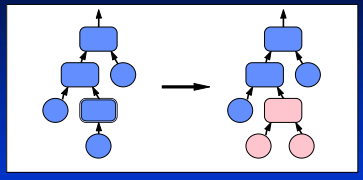
Terminals: numerical constants or input (x-) variables.
Nodes: mathematical operators or program functions.

GP/ GC FUNCTION (PARSE) TREE



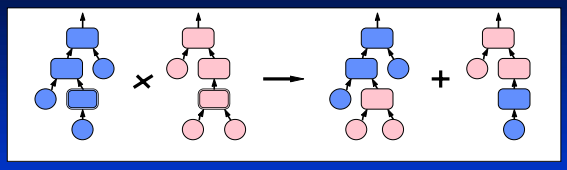
- **The genes are organized into a chromosome with a tree structure.**
  - The number of nodes is variable.
  - Nodes can be of any type.
- **To evaluate the tree, each node evaluates its argument nodes, processes the returned values, and returns its own value.**

GP MUTATION



- **Each node accepts and returns values of the same type.**
  - Trees are modular, allowing logically consistent changes to be introduced.
- **A node is randomly chosen and modified.**
  - It may be given a different operator with the same number of arguments.
  - It may be replaced by a new random sub-tree.
- **Terminals are mutated by slightly perturbing their numerical values, or randomly choosing a new input variable.**

GP CROSSOVER



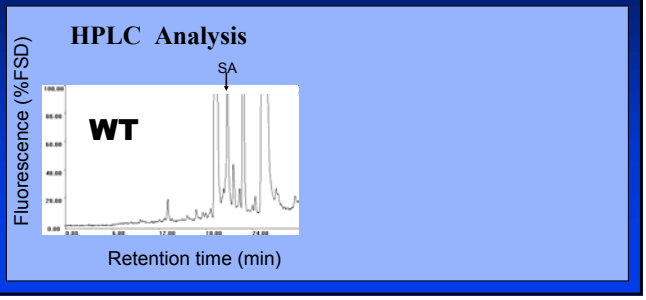
- Two parents are chosen with a probability proportional to their fitness.
- A node is randomly chosen on each parent tree.
- The selected sub-trees are swapped.
  - The new trees are still syntactically correct.
- The new individuals replace less fit members of the population.

Specific advantages of Genomic Computing

- Not all variables are used – this at once both (a) cuts the search space hugely and (b) makes the rules intelligible
- Evolutionary computing methods build on partially successesful rules and are highly efficient at negotiating complex search spaces.
- Preprocessing and normalization are unnecessary – the system learns what it is best to normalize to.
- Ranking of objects exploits the full range of information available (conventional methods throw it away), and ranking of variables forces explanation to be as simple as useful – which avoids overfitting and greatly improves generalisation

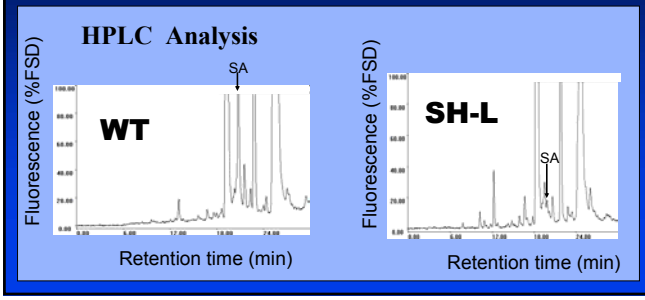
Salicylate experiment – effect of salicylate hydroxylase in plants following infection?

(data of Rob Darby & John Draper, Aberystwyth)



Salicylate experiment – effect of salicylate hydroxylase in plants following infection?

(data of Rob Darby & John Draper, Aberystwyth)



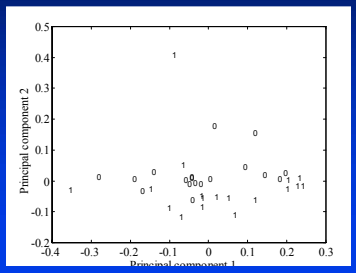
Deductive analysis of the salicylate experiment

- The plants containing SH-L are indeed more sensitive than is the WT to the subsequent wounding
- The salicylate concentration is indeed much lower

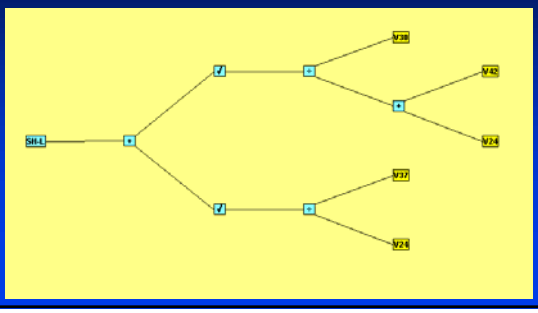
Deductive analysis of the salicylate experiment

- The plants containing SH-L are indeed more sensitive than is the WT to the subsequent wounding
- The salicylate concentration is indeed much lower
- This is (merely) consistent with its involvement in the normal defence response
- However, while the LC data show us the changes, they do not show us which changes matter for the problem of interest

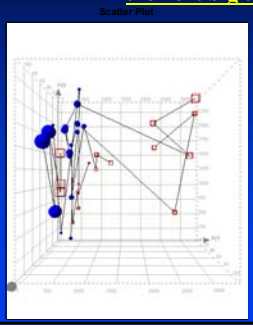
Unsupervised: PCA fails to discriminate plants with and without SH-L
(coded 1 / 0 in this long-time-course experiment)



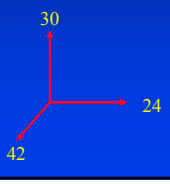
Supervised: using GP, the top rule that evolves is both simple and accurate (gets 95% of all samples correctly assigned)



A plot of the 'top 3' variables allows visualisation of what is important – closed circles contain SH-L – and light up 2 other important but previously neglected metabolites - the next big thing in plant defence research



NB the 3D plot also shows that the problem is not linearly separable

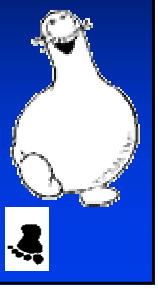


Metabolic footprinting – a novel genome-wide approach for functional genomics

- Metabolic 'fingerprinting' of microbes is difficult as quenching must be fast (turnover time ~ concentration/flux) and a small intracellular space must be separated from a large extracellular one
- So, we recognise that in unbalanced growth what microbes will excrete is a function of which pathways are operating
- We study therefore not the metabolic 'fingerprint' but the metabolic 'footprint' of what they leave behind in the medium

Metabolic footprinting experiments

- Use *Saccharomyces cerevisae*, because of its sequenced genome and the availability of a complete series of single-gene knockout strains
- Grow cultures in microtitre plates, take supernatant, mix with solvent and squirt directly into an electrospray mass spectrometer
- ('DIMS' – Direct Injection Mass Spectrometry)



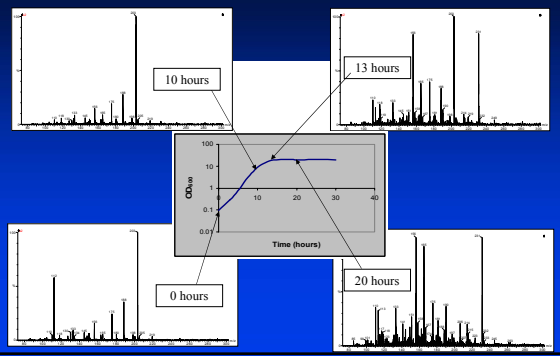
nature biotechnology

High-throughput classification of yeast mutants for functional genomics using metabolic footprinting

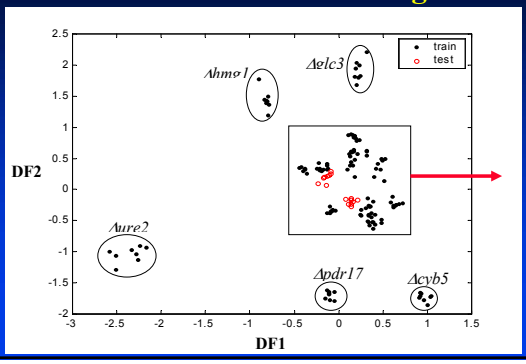
Jess Allen[1], Hazel M Davey[1], David Broadhurst[1], Jim K Heald[1], Jem J Rowland[2], Stephen G Oliver[3] & Douglas B Kell[1]

[1]Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, Aberystwyth SY23 3DD, UK. [2]Department of Computer Science, University of Wales, Aberystwyth, Aberystwyth SY23 3DB, UK. [3]School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Rd., Manchester M13 9PT, UK. [4]Present address (from 1 September 2003): Department of Chemistry, UMIST, Sackville St., P.O. Box 88, Manchester M60 1QD, UK. Correspondence should be addressed to D.B.K. (dbk@umist.ac.uk).

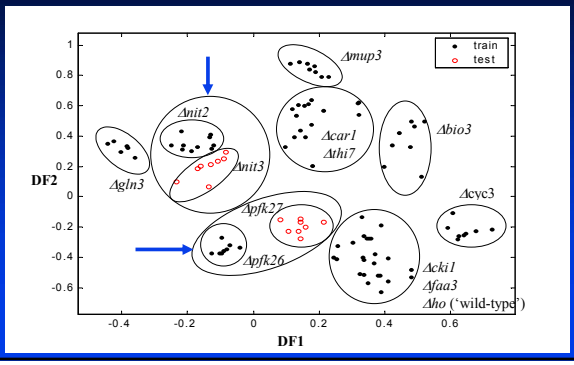
The metabolic footprint varies with growth phase and is richest in stationary phase



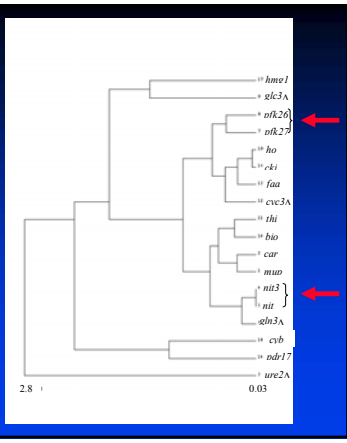
The metabolic footprints differentiate strains, and similar strains cluster together



Metabolic footprints differentiate strains, similar strains cluster together, and can be used to predict 'unknowns'



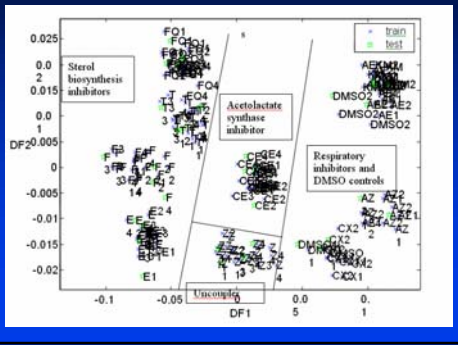
Hierarchical Cluster Analysis of strains using DFs shows natural groupings



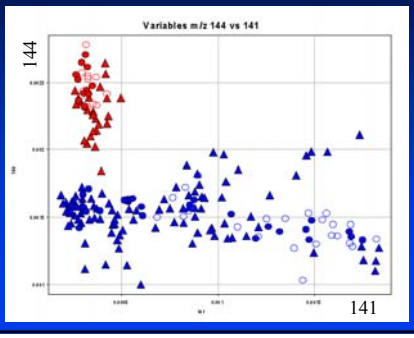
Genetic Programming can provide a simple rule for distinguishing strains:

- **IF *m/z*_201 > 0.00126 of TIC THEN deletant is 'nitrilase'**
- **This is true for both the training set and the unseen cross-validation and test sets**
- **By analysing m/z_201 using tandem mass spectrometry we should expect to identify the biochemical basis for the rule**

Mode of action may be discriminated, as well as effects of gene KOs



…and GP lights up 2 variables that discriminate uncouplers from the rest



Tuning mass spectrometers via genetic search

Anal. Chem. 2003, 75, 6679−6686

Explanatory Optimization of Protein Mass Spectrometry via Genetic Search

Seetharaman Vaidyanathan,†,‡ David I. Broadhurst,‡ Douglas B. Kell,*,†,‡ and Royston Goodacre†,‡

Department of Chemistry, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, U.K., and Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DD, U.K.

14 parameters to describe ESMS conditions

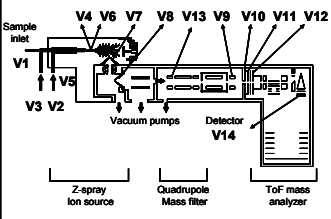


Figure 2

Table 1: Variables and their range.

Code	Variables	Range
V1	Sample flow rate (µL/min)	20-500
V2	Desolvation gas flow rate (L/h)	150-500
V3	Nebulizer gas flow rate (L/h)	10-20
V4	Source temperature (° C)	40-150
V5	Desolvation temperature (° C)	100-400
V6	Capillary voltage (V)	1500-3500
V7	Skimmer 1 (sample cone) voltage (V)	10-150
V8	Skimmer 2 (extraction cone) voltage (V)	0-10
V9	Transport hexapole voltage (V)	0-20
V10	Differential pumping Aperture voltage (V)	0-20
V11	Acceleration lens voltage (V)	0-200
V12	Focus voltage (V)	0-200
V13	Pre-filter voltage (V)	5-15
V14	MCP detector voltage (V)	2300-2700

5 proteins, each flying with very different efficiencies

(A) insulin (5.7kDa), (B) ubiquitin (8.6 kDa), (C) cytochrome *c* (12.3 kDa), (D) lysozyme (14.3 kDa), and (E) myoglobin (16.9 kDa), (F) a spectrum of an equimolar mixture of the five proteins, and (G) their combined theoretical mixture spectrum, all obtained under one set of instrumental conditions.

Note the huge difference between 'theoretical' and 'experimental' spectra in the mixtures (G vs F)

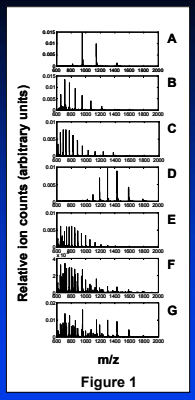
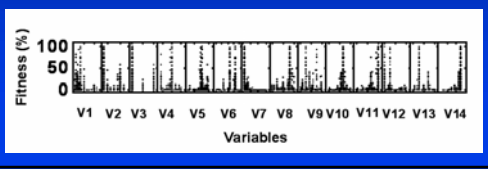


Figure 1

We used 6 generations, 60-80 runs in each generation, 439 experiments

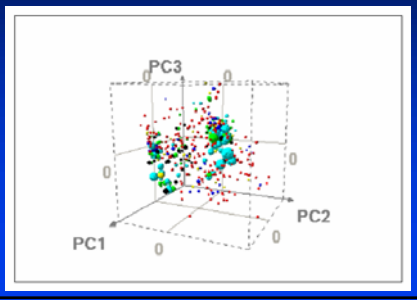
Fitness is a composite of TIC, equality of contribution of each protein and 'coverage' of charge states in mixture

- **Distribution of the relative fitness (%) for the 14 univariate variables (V1-V14) (data are the experimental results)**



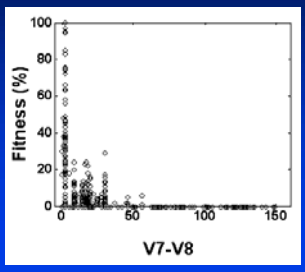
PCA analysis of the search space, showing strong multimodality/ epistasis

Size = fitness, Colour = generation number

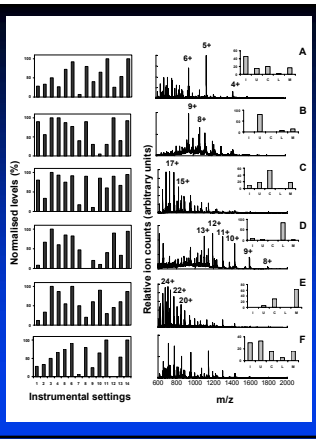


GP is used to find a rule that best explains fitness; many rules come up with an unusual combination of variables….

As well as low values of V7 (sample cone voltage) we want V8 (extraction cone voltage) to be just higher than it. A new relationship – from the data alone (no hypothesis!)



One may also evolve conditions in which only a particular protein is observed, although all 5 are present together



Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

[1]*Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK*
[2]*School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK*
[3]*Department of Computing, Imperial College, London SW7 2AZ, UK*
[4]*Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK*
[5]*School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK*

NATURE | VOL 427 | 15 JANUARY 2004 | 247

Robot scientist catches the world's imagination

A 'robot scientist' capable of generating hypotheses on the function of genes in yeast, with the ability to both design experiments and interpret the data produced has been developed by team of British scientists, led by Professor Ross King from the University of Wales, Aberystwyth. The research was published in Nature in January 2004, and the response from the world's media was extraordinary.

"The story was covered extensively in the UK and North America," says BBSRC media officer Andrew McLaughlin. "There was also wide coverage across Western Europe, the Middle East, Asia and Australia."

The research group, funded by BBSRC and EPSRC, the Wellcome Trust and PharmDM set the robot the problem of discovering the function of different genes in the baker's yeast *Saccharomyces cerevisiae*. The functions of about 30% of the 6,000 genes remains unknown, and with many of these genes thought to be common with those found in the human genome, such research could prove medically significant in the future. Although the problems set for the robot were relatively simple, the research addresses the increasing demand for automation in the biological sciences - the potential is now there to solve real-world problems.

Professor Ross King with the celebrated robot.

And in case you missed it... the BBSRC media release is at: www.bbsrc.ac.uk/media/pressreleases/04_01_14_rob.html

The Robot Scientist

- Some background knowledge (metabolic pathways are graphs (i.e. networks) with nodes and edges)
- Aim is to find the site of a genetic lesion (an edge in the graph) on the basis of growth experiments (organisms whose graphs are unconnected to the synthesis of essential nutrient sources do not grow in their absence)
- Chooses a growth experiment to perform and then does it
- Iterates around the cycle until it has a final hypothesis

Logical model of amino acid metabolism

Robot scientist abduction of facts

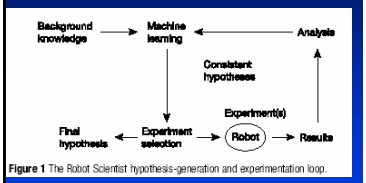
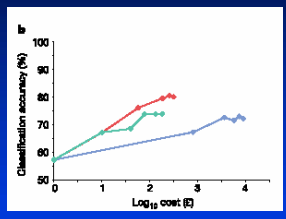


Figure 1 The Robot Scientist hypothesis-generation and experimentation loop.



Robot Scientist, random, naïve (cheapest expt)

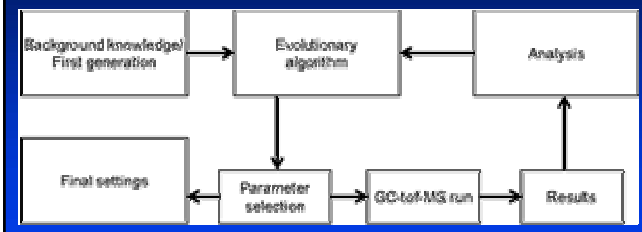
Anal. Chem. **2005**, *77*, 290−303

Closed-Loop, Multiobjective Optimization of Analytical Instrumentation: Gas Chromatography/ Time-of-Flight Mass Spectrometry of the Metabolomes of Human Serum and of Yeast Fermentations

Steve O'Hagan, Warwick B. Dunn, Marie Brown, Joshua D. Knowles, and Douglas B. Kell*

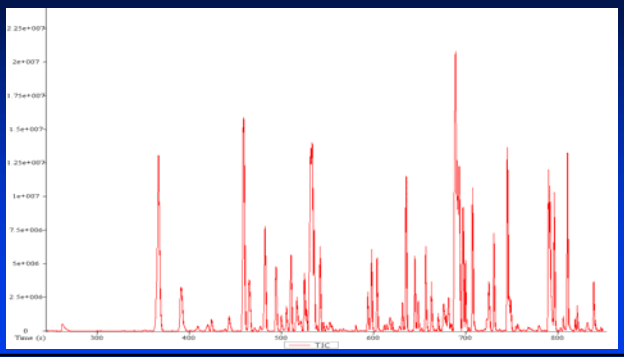
School of Chemistry, University of Manchester, Faraday Building, Sackville Street, P.O. Box 88, Manchester M60 1QD, U.K.

Closed-loop, multi-objective optimisation of GC-tof analysis

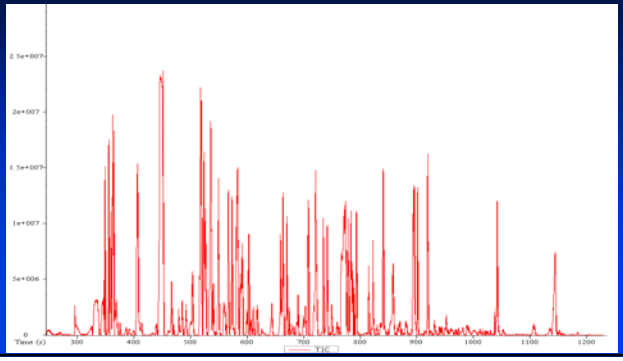


Size of search space > 200,000,000 combinations of settings….

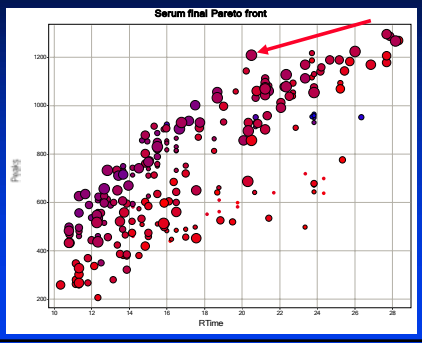
Serum, generation 1



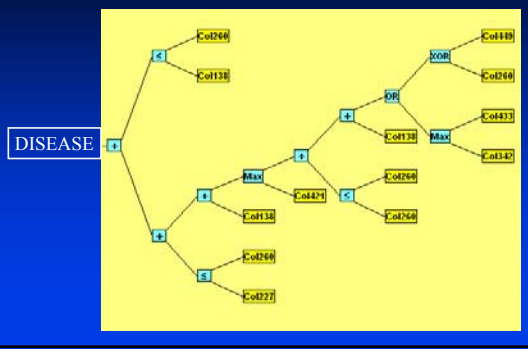
Serum, final



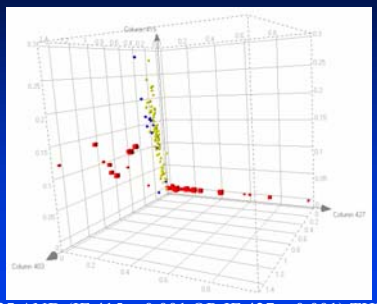
Closed-loop, multi-objective optimisation of GC-tof analysis



Disease diagnosis by GC-tof- MS – look for metabolites that discriminate 'cases' from 'controls'



Disease diagnosis by GC-tof- MS – look for metabolites that discriminate 'cases' from 'controls'



IF 403 < 0.035 AND (IF 415 < 0.001 OR IF 427 > 0.001) THEN disease
IF 403 > 0.015 AND (IF 415 <0.012) THEN disease

NFκB (1)

- NF-κB is a nuclear transcription factor that can modify the expression of many other genes
- It is held inactive in the cytoplasm of non-stimulated cell by three IκB isoforms.
- It is widely and diversely implicated in cancer, apoptosis and in diseases such as arthritis

Evidence for 'involvement'

1. c-rel (an alternative NF- κB subunit like p65) and Bcl-3 (an IκB) are known oncogenes.
2. Increased nuclear localisation of NF- κB is associated with many cancers.
3. NF-κB clearly regulates both cell cycle and apoptosis and is involved in the response to DNA damage.

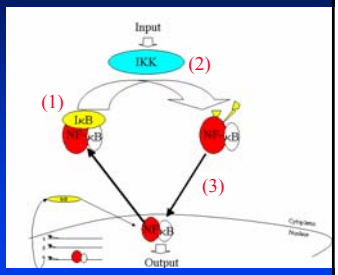
Question 1: so what is a good drug target in the NFκB pathway?
Question 2: and how do we measure that?

The big question… (aka the 'crosstalk problem')

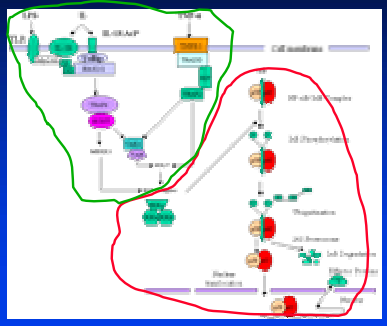
How can the same thing (i.e. NF-κB) – it is assumed by changes in its concentration in the nucleus – be 'involved' both in cell proliferation in cancer and in apoptotic cell death (two processes that are pretty well opposite in character)?!

Summary of NF-κB – 3 steps

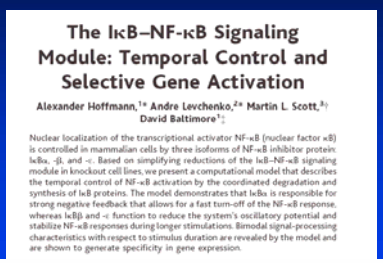
1. NF-κB is a nuclear transcription factor and is held inactive in the cytoplasm of non-stimulated cell by three IκB isoforms
2. During cell stimulation, the IKK complex is activated, leading to phosphorylation and ubiquitination (and removal) of the IκB proteins.
3. Free NF-κB translocates to the Nucleus, activating genes including IκBα. IκBβ& -ε are synthesised at a steady rate, allowing for complex temporal control of NF-κB activation involving negative feedback



Many effectors (e.g. TNFα) can activate IKK



Hoffman *et al* (2002) produced a reduced model for cells lacking two IκB isoforms (IκBβ and IκBε)

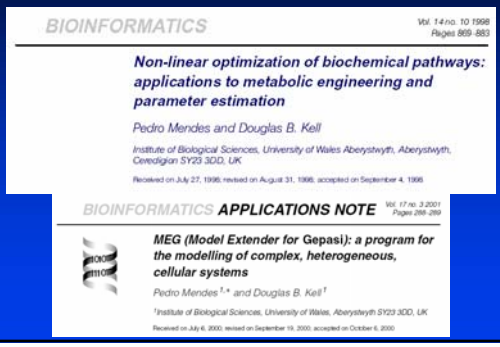


The IκB–NF-κB Signaling Module: Temporal Control and Selective Gene Activation

Alexander Hoffmann,[1]* Andre Levchenko,[2]* Martin L. Scott,[3] David Baltimore[1]†

Nuclear localization of the transcriptional activator NF-κB (nuclear factor κB) is controlled in mammalian cells by three isoforms of NF-κB inhibitor protein: IκBα, -β, and -ε. Based on simplifying reductions of the IκB–NF-κB signaling module in knockout cell lines, we present a computational model that describes the temporal control of NF-κB activation by the coordinated degradation and synthesis of IκB proteins. The model demonstrates that IκBα is responsible for strong negative feedback that allows for a fast turn-off of the NF-κB response, whereas IκBβ and -ε function to reduce the system's oscillatory potential and stabilize NF-κB responses during longer stimulations. Bimodal signal-processing characteristics with respect to stimulus duration are revealed by the model and are shown to generate specificity in gene expression.

Hoffman et al used the modelling system *Gepasi* written by Pedro Mendes



BIOINFORMATICS Vol. 14 no. 10 1998 Pages 869–883

Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation

Pedro Mendes and Douglas B. Kell

Institute of Biological Sciences, University of Wales Aberystwyth, Aberystwyth, Ceredigion SY23 3DD, UK

Received on July 27, 1998; revised on August 31, 1998; accepted on September 4, 1998

BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 3 2001 Pages 288–289

MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous, cellular systems

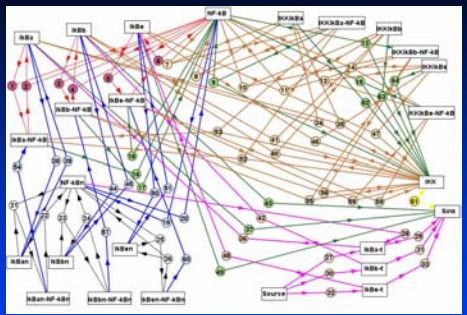
Pedro Mendes[1,*] and Douglas B. Kell[1]

[1]Institute of Biological Sciences, University of Wales, Aberystwyth SY23 3DD, UK

Received on July 6, 2000; revised on September 19, 2000; accepted on October 6, 2000

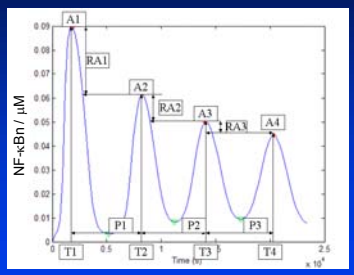
We have reproduced this model (modified to remove mistakes in the original, now corrected) using Gepasi

The model has 64 unidirectional reactions & 26 variables

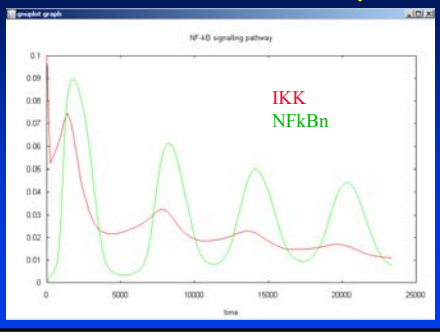


Violet red circles = IκB-NF-κB cytoplasmic reactions; Blue Arrows and circles = Nuclear Transport; Magenta Arrows and Pink circles = IκB mRNA synthesis (including transcription, translation and degradation); Black Arrows and white circles = IκB-NF-κB nuclear reactions; Light Green Arrows and circles = IκB Phosphorylation and Degradation reactions; Brown Arrows and brown circles = Bimolecular IKK- IκB and tri-molecular IKK- IκB-NF-κB; Yellow Arrows and circles = IKK slow adaptation coefficient

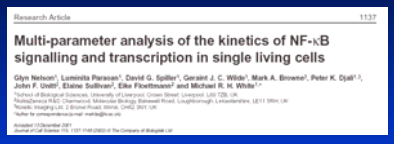
Cartoon of nuclear NF-κB after IKK addition



After pre-equilibration for 2000s, IKK is 'added' at 0.1 μM

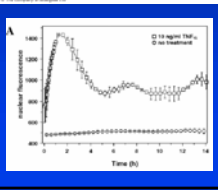


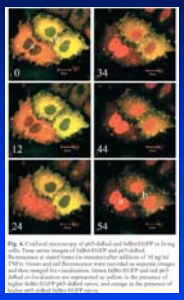
"Real" oscillations of GFP-NFκBn observed microscopically (and averaged)


Research Article 1137

Multi-parameter analysis of the kinetics of NF-κB signalling and transcription in single living cells

Glyn Nelson[1], Luminita Paraoan[1], David G. Spiller[1], Geraint J. C. Wilde[1], Mark A. Browne[1], Peter K. Djali[1,2], John F. Unitt[2], Elaine Sullivan[2], Eike Floettmann[2] and Michael R. H. White[1,*]



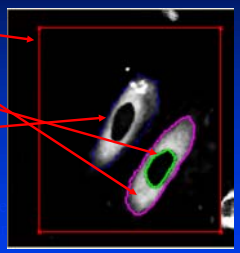


Tracker

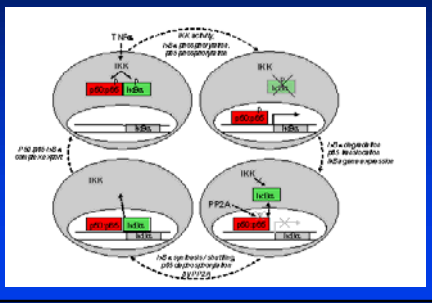
- **A work bench with tracking procedures at three levels**
  - Manual
  - Semi-automatic
  - Automatic
- **Strong browsing capabilities with result display and export.**

Four regions

- ROI
- Cell boundary
- Nuclear boundary
- User input

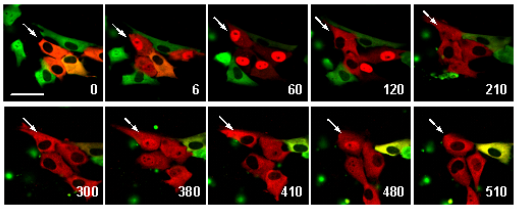


"Real" oscillations of GFP-NFκBn observed microscopically with labelled IκBα and NFκB



Nelson et al 2004

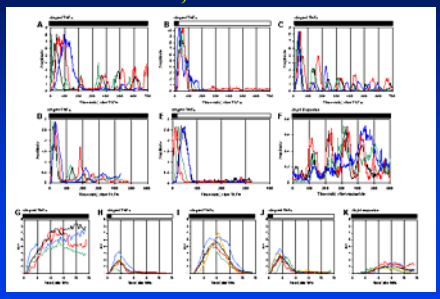
"Real" oscillations of GFP-NFκBn observed microscopically with labelled IκBα and NFκB



Nelson et al 2004

NB we measure individual cells, not ensembles

The timing and amount of oscillations depend strongly on the type of stimulation (various amounts and times of TNFα, different individual cells)



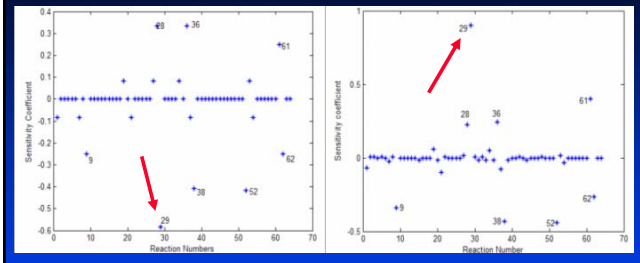
Nelson et al 2004

What about the model? Sensitivity analysis

- A generalised form of the control coefficients of MCA
- Dimensionless
- Describe quantitatively which reactions are most 'important'
- In favourable cases (especially steady states) there are summation theorems

$$S_P^M = \frac{\delta M / M}{\delta P / P}$$

Sensitivity coefficients of T3 for δP of 10% or 100%



- Only 8 reactions have significant sensitivity coefficients when T3 is measured
- Note the change in sign for reaction 29 – very nonlinear system

When all the outputs for NFkBn fluctuations are considered, only 8-9 out of the 64 reactions show any significant sensitivity to their parameters

9 important reactions

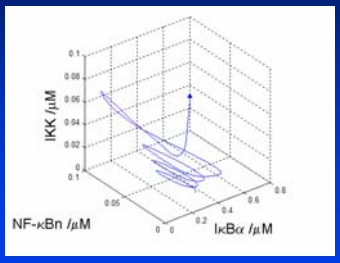
9: IKKIκBα-NF-κB catalytic rate constant
28: IκBα (IκBα-t) Inducible mRNA synthesis rate constant
29: IκBα (IκBα-t) mRNA degradation rate constant
34 : IKKIκBα association rate constant
36: Constitutive IκBα translation rate constant
38: IκBαn nuclear Import Rate constant
52: IKKIκBα-NF-κB association rate constant
61: IKK signal onset slow adaptation coefficient
62: IKKIκBα catalysis rate constant

What do they have in common?

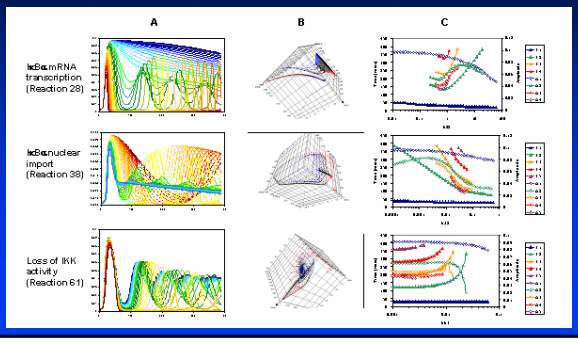
They all involve free IKK and/or IκBα

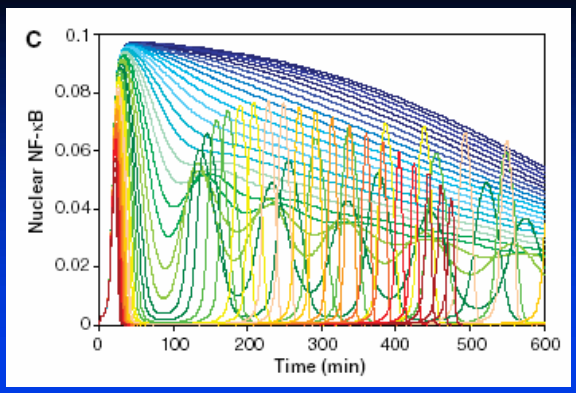
9: IKKIκBα-NF-κB catalytic rate constant
→ *28*: IκBα (IκBα-t) Inducible mRNA synthesis rate constant
29: IκBα (IκBα-t) mRNA degradation rate constant
34 : IKKIκBα association rate constant
36: Constitutive IκBα translation rate constant
→ *38*: IκBαn nuclear Import Rate constant
52: IKKIκBα-NF-κB association rate constant
→ *61*: IKK signal onset slow adaptation coefficient
62: IKKIκBα catalysis rate constant

A phase plane plot shows the intimate connection between IKK, IκBα and NFκBn

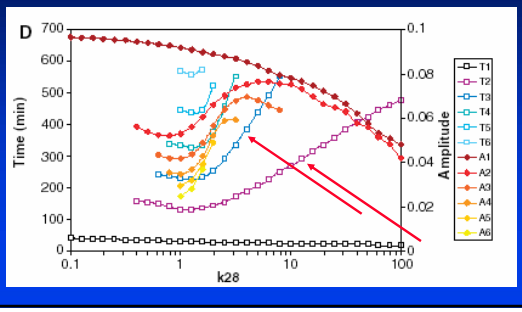


Effect of changing 3 parameters by ± 2 logs on NFκBn dynamics

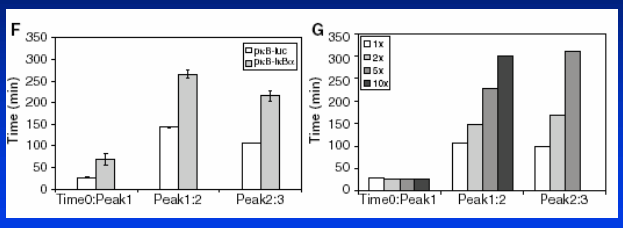




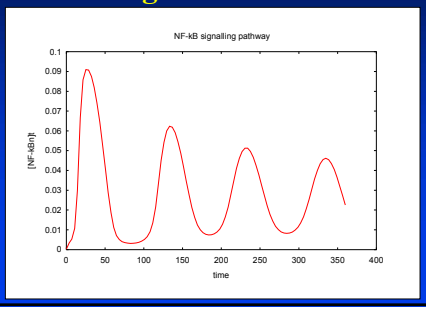
Prediction: increasing k28 will increase the period of the oscillations (e.g. T2 and T3)



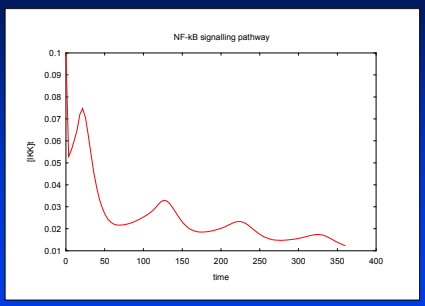
Experiment (left) matches simulation (right)



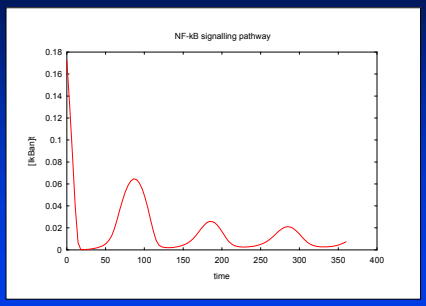
Mathematical challenge is the Inverse Problem – work out the system that gave THIS time series



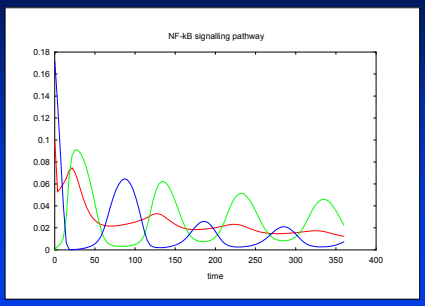
… and this time series



… and this time series

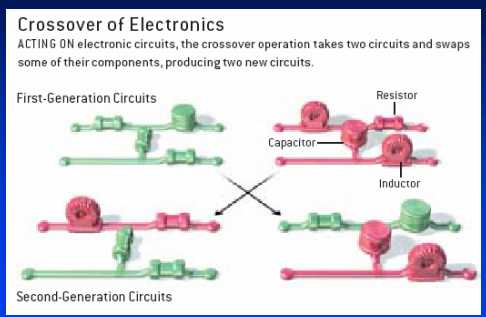


…and all 3 (or 23) together

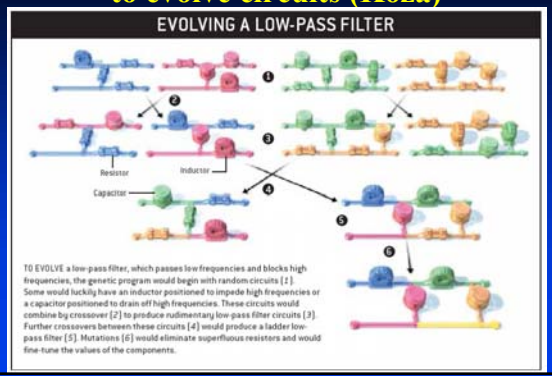


One example of solving inverse problems using genetic programming (Koza)

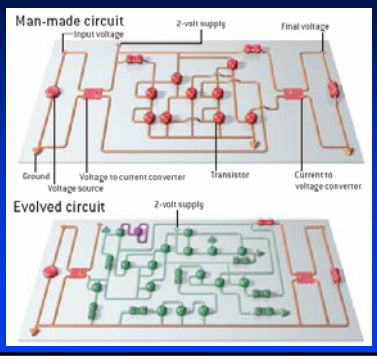
Use of genetic programming to evolve circuits (Koza)



One example – genetic programming to evolve circuits (Koza)



Circuits as functioning i/o systems



We usually consider biological circuit elements such as enzymes as 'responding' solely to amplitudes

e.g. Michaelis-Menten:

$$v = (V_{max}.S)/ (S + K_m)$$

Thus, *v* depends ONLY on the 'instantaneous' concentration of S

What is it in these hugely complex dynamics that actually controls downstream events, including cell fate?

If we had set up an assay that recorded solely the (change in)the amplitude or concentration of NF-κB at time t we would have been completely misled as to any possible efficacy of a drug, as the encoding of the important signal is not simply in the concentration but the frequency.

Frequency encoding

- Having the effective signal frequency-encoded allows the same 'medium' (NF-kB) to carry different 'messages' using changes in the frequency or dynamics rather than the amplitude of oscillatory signals *per se*
- There is thus no 'crosstalk' (and no crosstalk problem)
- But this also means that great care must be used if such systems are to be exploited for providing novel drug targets simply by inhibiting particular steps, as the downstream events are not easily related to the activities of the individual steps
- (Additional means of avoiding crosstalk are likely also present, e.g. extra transcription factors providing a logical AND.)
- More generally, we need to recognise signalling systems as signal processing systems

Signalling as signal processing

- **This is not just semantics; in the signal processing view of signalling we thus lay stress not so much on the amounts and nature of signalling molecules but on their dynamics and on the processing structures that are necessary to distinguish different dynamics – just as we might distinguish 2 speakers via their voice patterns although both use the same intermediary medium**

Any metabolite or signalling pool serves to act as a low-pass filter as it must be 'filled up' before its concentration is large enough to be converted to the next product

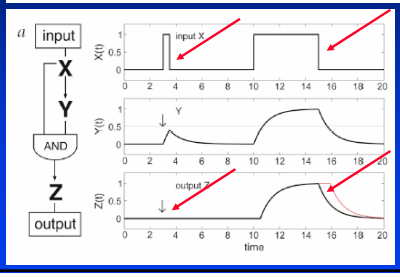
A → B → C → D → E
e1 e2 e3 e4

The pool size is like a capacitor and the kinetics of production like a resistor

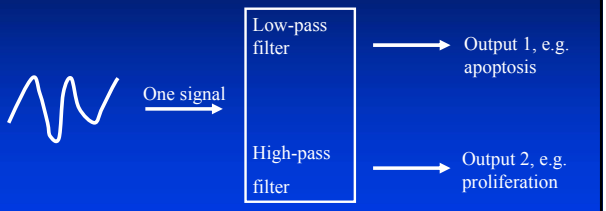
Other network motifs are emerging, e.g. the coherent feedforward loop

Network motifs in the transcriptional regulation network of *Escherichia coli*

Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

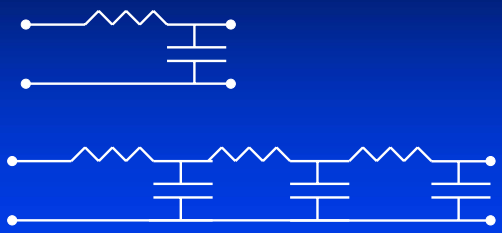


The same signal can lead to two different outputs depending on the filtering/detector



Multiple LP filters in series act as a delay line

This may in part explain why pathways have multiple steps (although amplification can be an important reason too)



Sensitivity analysis is generally useful for looking at complex systems (e.g. in a study of drug target identification)

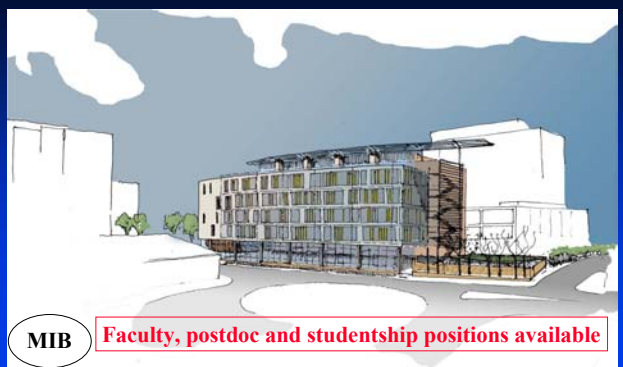
Comparative and Functional Genomics
Comp Funct Genom 2004; **5**: 304–327.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/cfg.411

Research Article

Comparative genomic assessment of novel broad-spectrum targets for antibacterial drugs

Thomas A. White[1] and Douglas B. Kell[2]*
[1]Department of Biology, University of York, Box 373, Heslington, York YO10 5YW, UK
[2]Department of Chemistry, UMIST, Faraday Building, Sackville St, PO Box 88, Manchester M60 1QD, UK

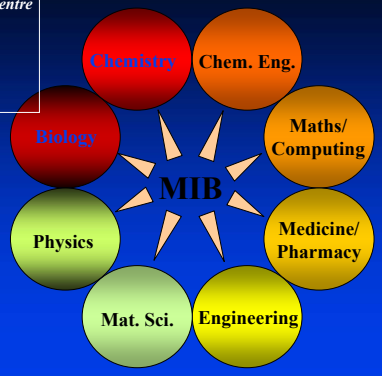
The Manchester Interdisciplinary Biocentre (2005-)



MIB

Faculty, postdoc and studentship positions available

The MIB model
- a major international research centre
- a partner to the disciplines
- parallel to schools/faculties
- one of a number of expected IRCs



http://www.mib.ac.uk

MIB

The MIB – level D

Overall conclusions

- Systems biology represents and requires a judicious interplay between high-quality, large-scale experimentation and suitable computational modelling.
- This is of course a very multi- and interdisciplinary endeavour, and Collaborative Systems Biology endeavours (localised or distributed) may be a useful strategy
- Technology development is a major area; this is not, and certainly not only, hypothesis-dependent science
- Vertical integration and studying systems as systems is crucial
- Systems and subsystems talk to each other in complex, nonlinear ways – this is the 'language of cells' that we must learn to understand

Credits

Aberystwyth

Jem Rowland
Ross King
Royston Goodacre
Mike Winson
David Broadhurst
Hazel Davey
Jess Allen

Manchester

Royston Goodacre
Josh Knowles
Rick Dunn
Dave Ellis
Steve O'Hagan
Adaoha Ihekwaba
Marie Brown
Dave Broomhead
Hailin Shen
Irena Spasic
Seetharaman Vaidyanathan

Steve Oliver
Phil Baker
Louise Kenny
Andy Hayes
Nikki Burton

Liverpool: Mike White, Dave Spiller. **Pfizer:** Neil Benson, Rachel Grimley

Another kind of nonlinear, oscillatory dynamics…..the music of Dr Subhendu Ghosh

- The auspicious day has come….it is the time to pack up the pending work…it is the time to welcome you…..

Metabolomics, machine learning and modelling: towards an understanding of the language of cells

Douglas Kell
School of Chemistry, University of Manchester,
MANCHESTER M60 1QD, U.K.
dbk@man.ac.uk
http://dbk.ch.umist.ac.uk/
http://www.mib.ac.uk

MANCHESTER 1824 The University of Manchester

Aber Genomic Computing