Towards Systemic Views of Gene Expression and Cell Metabolism

Heikki Hyötyniemi
HUT / Control Engineering

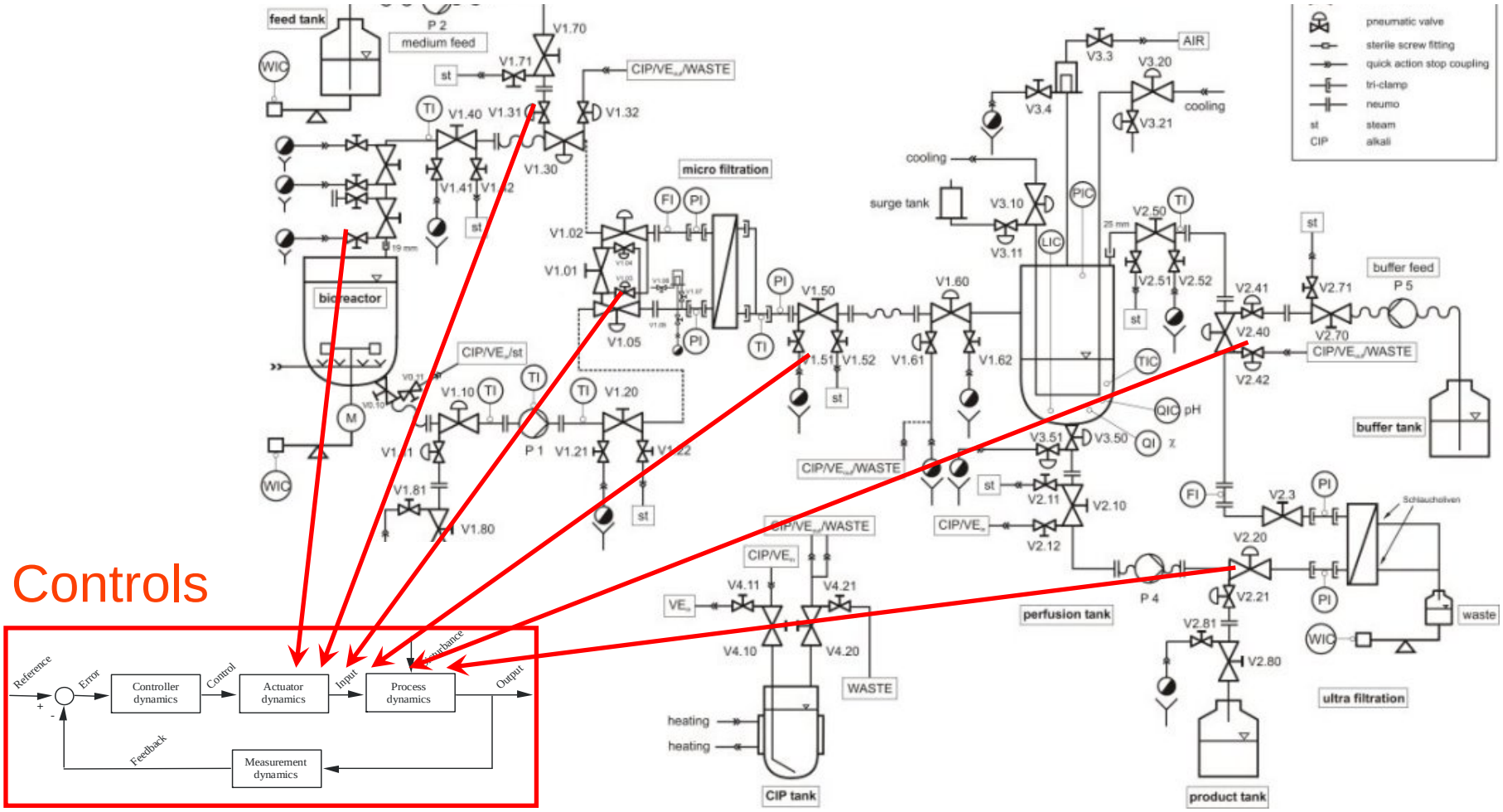
Presentation at OSSiBE3 – NeoBio Seminar,
June 9, 2005

What is our competence?

HELSINKI UNIVERSITY OF TECHNOLOGY
Control Engineering Laboratory
Cybernetics Group

Industrial systems becoming ”metabolic cells”

Same dilemma: How to understand complexity?



Controls

- Contribution of systems engineering:
  - ”Holistic understanding”, seeing complex systems in a perspective (?)
  - Understanding the relevance of dynamic phenomena
  - Conceptual tools – appropriate mathematical models and methods
  - Analogues to exploit
- However, traditional control engineering is still extremely *reductionistic* – just as today’s biology!?
- Contents of the speech:
  - Find a generic, *truly systemic* representation of cell phenomena
  - Find ways to exploit that representation
  - ... And present preliminary results

Cybernetics Rules!

... But what are those rules?

Cybernetics Group

Let us find it out!

http://www.control.hut.fi/cybernetics

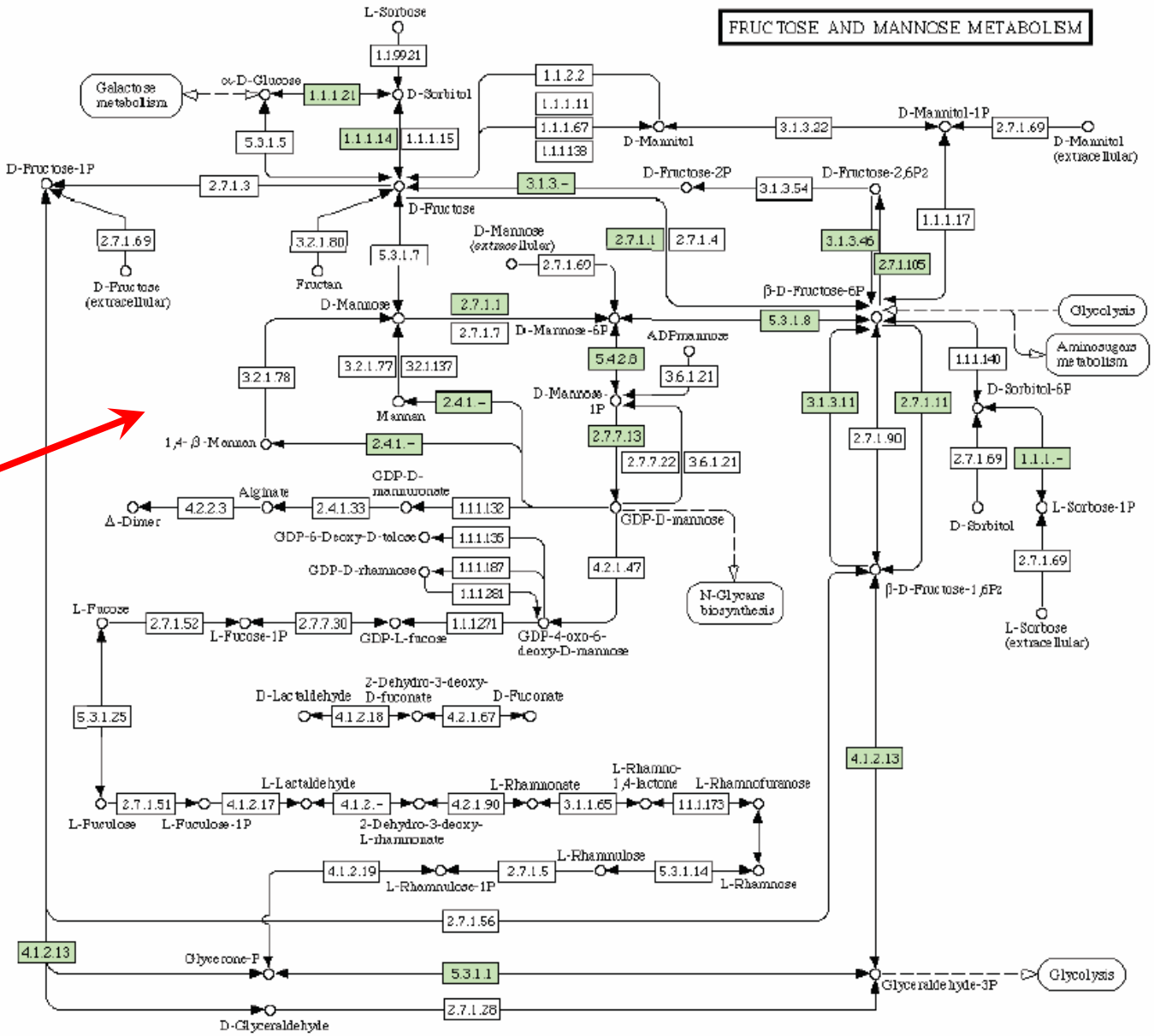
HELSINKI UNIVERSITY OF TECHNOLOGY
Control Engineering Laboratory
Cybernetics Group

Neocybernetic starting points

- The details are abstracted away, holistic view from the above is applied
- There exist local actions only, there are no structures of centralized control
- It is assumed that the underlying interactions and feedbacks are consistent, maintaining the system integrity
- This means that one can assume *stationarity* and *dynamic balance* in the system in varying environmental conditions
- An additional assumption: Linearity is pursued as long as it is reasonable

What are the consequences concerning NETWORKS?

- Chemical processes can be *very* complex ...
- Metabolic sub-network in yeast
- However, not *so* simple: There exist overlapping sub-networks

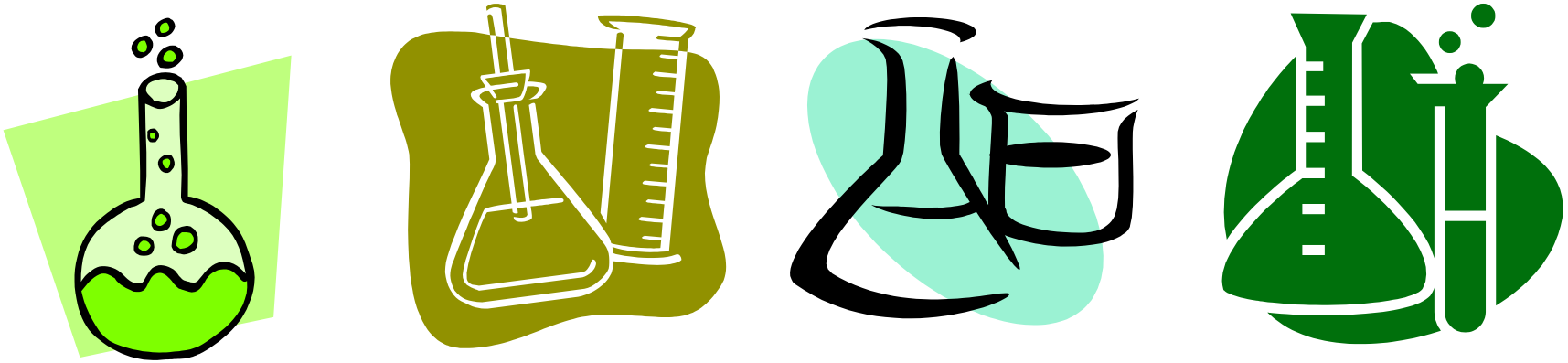


Chemical systems: Starting from bottom

- Forget the metabolic networks for a moment –
  Can one define a ”cybernetic model” for chemical systems?
- Prototypical reaction

$$a_1\mathrm{A}_1 + \cdots + a_N\mathrm{A}_N \longrightarrow b_1\mathrm{B}_1 + \cdots + b_M\mathrm{B}_M\,, \qquad \Delta H$$

- How to ”cybernetize” such models?



Intuition #1: Problem formulation

- First augment the reaction:

Enthalpy

$$a_1 C_1 + \cdots + a_m C_m \quad \underset{k_b}{\overset{k_f}{\rightleftarrows}} \quad b_1 C_1 + \cdots + b_m C_m, \qquad \Delta H$$

Here, there are all chemicals on both sides; a_i and b_j can be zeros. Reactions are assumed reversible.

- Collect all chemical concentrations in a single data matrix u; then one can write $\Delta u = r\,\Theta$ where r is *reaction rate*, and

$$\Delta u = \begin{pmatrix} \Delta C_1 \\ \vdots \\ \Delta C_m \\ \hline \Delta T \end{pmatrix} \qquad \text{and} \qquad \Theta = \begin{pmatrix} b_1 - a_1 \\ \vdots \\ b_m - a_m \\ \hline c_T \end{pmatrix}$$

- If there are many simultaneous reactions, the changes in the system state can be expressed in the matrix form

$$\Delta u = r^T \Theta$$

- This kind of approach is known as ”flux balance analysis” (also compare to *reaction invariants*)
- However, it is difficult to keep track of all fluxes (for example, to master temperatures, the system should be isolated)
- Flux balance captures the *stoichiometric balance* = more or less *formal balance*
- There is no information of whether the reactions actually take place or not – one needs the functional or dynamic balance

Intuition #2: Thermodynamic equilibrium

- Reaction speed k_f is related to probability of unit reaction is related to probability of the constituents to be located near enough each other is related to chemical *concentrations*
- In strong liquids *activities* substitute concentrations
- Reaction speed is also dependent of the temperature (Arrhenius law) – altogether

$$k_\mathrm{f} = c_\mathrm{f}\ e^{-a_T/T}\ C_1^{a_1} \cdot \cdots \cdot C_n^{a_n} \qquad k_\mathrm{b} = c_\mathrm{b}\ e^{-b_T/T}\ C_1^{b_1} \cdot \cdots \cdot C_n^{b_n}$$

- In dynamic (cybernetic, homeostatic) equilibrium, the speeds of forward and backward reactions are equal, and exists

$$K = \frac{e^{-b_T/T}}{e^{-a_T/T}} \frac{C_1^{b_1} \cdot \cdots \cdot C_n^{b_n}}{C_1^{a_1} \cdot \cdots \cdot C_n^{a_n}}$$

Intuition #3: Linearity

- The function is purely multiplicative – take logarithms:

$$\log K = \left(a_T - b_T\right)1/T + \left(b_1 - a_1\right)\log C_1 + \cdots + \left(b_n - a_n\right)\log C_n$$

- To get rid of constants and logarithms, it is also possible to differentiate the expression

$$0 = \left(b_T - a_T\right)\Delta\left(\frac{1}{T}\right) + \left(b_1 - a_1\right)\frac{\Delta C_1}{\overline{C}_1} + \cdots + \left(b_n - a_n\right)\frac{\Delta C_n}{\overline{C}_n}$$

  where the variables are deviations from the nominal values, divided by those nominal values = relative changes

- The differentiated model is only locally applicable, valid in the vicinity of the nominal value

- Acidity is logarithmic measure, and its absolute value can be included in data:

$$\mathrm{pH} = -\lg C_{\mathrm{H}^+}$$

- Non-balance compounds can be included in data: Assume that G denotes the rate of change, or flow, into / out from the system, so that in balance, for example

$$\frac{\Delta \dot{C}_0}{\bar{\dot{C}}_0} = b_T \Delta\left(\frac{1}{T}\right) + b_1 \frac{\Delta C_1}{\bar{C}_1} + \cdots + b_n \frac{\Delta C_n}{\bar{C}_n}$$

$$u' = \begin{pmatrix} 1/T \\ \hline \mathrm{pH} \\ \hline \log C_1 \\ \vdots \\ \log C_n \\ \hline \log G_1 \\ \vdots \\ \log G_1 \end{pmatrix} \qquad u = \begin{pmatrix} \Delta\left(\frac{1}{T}\right) \\ \hline \Delta\mathrm{pH} \\ \hline \frac{\Delta C_1}{\bar{C}_1} \\ \vdots \\ \frac{\Delta C_n}{\bar{C}_n} \\ \hline \frac{\Delta G_1}{\bar{G}_1} \\ \vdots \\ \frac{\Delta G_m}{\bar{G}_m} \end{pmatrix}$$

Relative change in flow

Intuition #4: Multiple reactions

- Now, when the reaction parameters are collected in vector θ, there holds

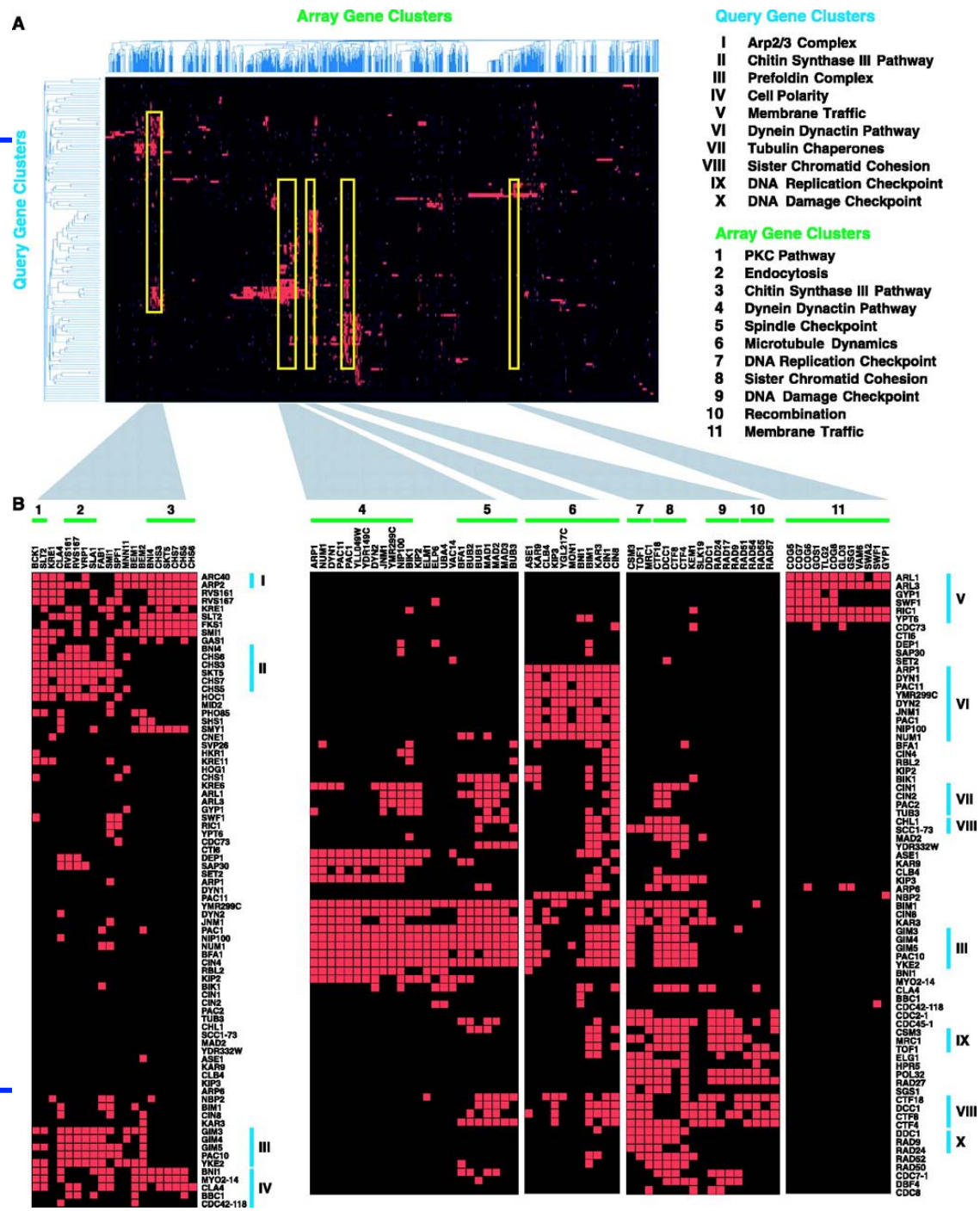
  $$0 = \theta^T \; u$$

  This holds also if there exist *simultaneous reactions*, so that θ is a matrix
- *Compare to flux balance analysis*: Now one only needs to study *levels*, not *changes*
- This is essential in complex chemical systems: The levels can better be controlled than the individual reactions
- However, the genetic system is esentially a part of the metabolic system, controlling it – what can be said about *it*?

- Example of gene interactions
- Rows: 132 query genes

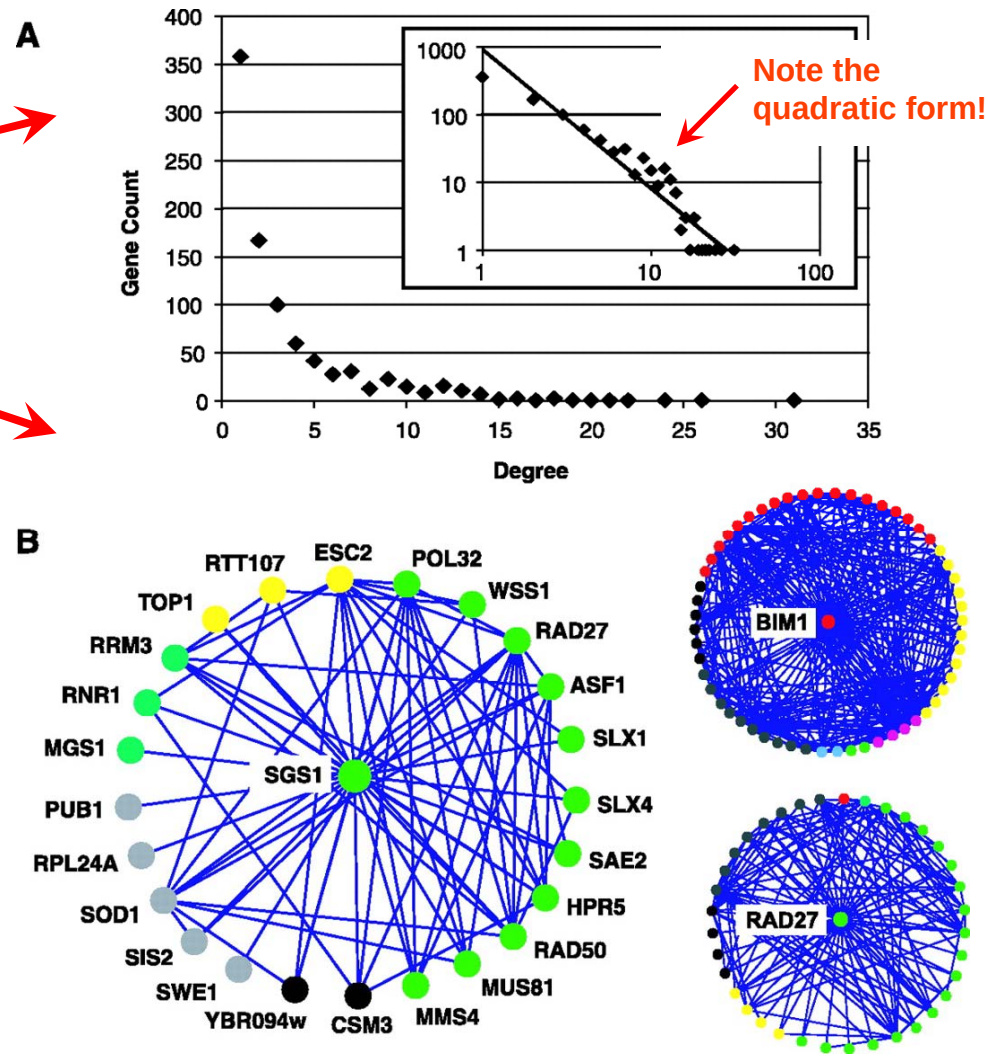
  Columns: 1007 array genes
- *Cluster trees* used to organize genes to show similarities

Science, Vol 303, Issue 5659, 808–813, February 6, 2004
50 authors!



HELSINKI UNIVERSITY OF TECHNOLOGY
Control Engineering Laboratory
Cybernetics Group

- Above, the gene connectivities shown
- Below, the topology of the genetic network around three query genes (named *SGS1*, *RAD27*, and *BIM1*)
- Clearly, such simple projections do not reveal the structure of the whole network



Approaches to complex networks

- Graph theory
  - Connections between nodes are ”crisp”
  - However, there is a continuum of interaction effects: The connections are not of ”all-or-nothing” type
- Bayesian networks
  - Strong probabilistic theory – *assuming that assumptions hold* ...
  - However, the ”nodes” in real networks are often not independent of each other: Loops and alternative paths exist in complex networks
- Now: **Neocybernetic framework**
  - Numeric, non-crisp connections, **fully connected**
  - **”Pancausality”** taken as the starting point: It is assumed that, after all, **all** nodes are causes and **all** are effects – opposite approach!

Cybernetic intuition #1: Stationarity & statistics

- Abstract away individual actions and realizations of interactions in the network
- Assume that the stationary state has been reached
- What are the statistical properties of the system?

- As advertised by *Albert-Laszlo Barabasi*, the emergent phenomena in networks are characterized by *power law*

$$y = z^D$$

- This dependency seems to govern all structures with *fractal* and *self-organized* structure
- This is taken as starting point here – and extended.

Cybernetic intuition #2: Multivariate nature

- Assume there are many variables of power law behavior:

$$\begin{cases} y = c_1 z_1^{D_1} \\ \vdots \\ y = c_n z_n^{D_n} \end{cases}$$

Parameter c_1 constant with respect to z_1

These can be combined:

$$\frac{y}{\bar{y}} = \left(\frac{z_1}{\bar{z}_1}\right)^{D_1} \cdots \left(\frac{z_n}{\bar{z}_n}\right)^{D_n}$$

- Further, there can exist various such dependencies
- Variables can be rearranged; assume there are (normalized) variables μ available:

$$\begin{cases} 1 = \mu_1^{\theta_{11}} \cdots \mu_m^{\theta_{1m}} \\ \vdots \\ 1 = \mu_1^{\theta_{n1}} \cdots \mu_m^{\theta_{nm}} \end{cases}$$

Cybernetic intuition #3: Linearity pursuit

- The same dependencies can be expressed in various ways; the equivalent static set of equations (after taking logarithms) is

$$\begin{cases} 0 = \theta_{11} \log \mu_1 + \cdots + \theta_{1m} \log \mu_m \\ \quad \vdots \\ 0 = \theta_{n1} \log \mu_1 + \cdots + \theta_{nm} \log \mu_m \end{cases}$$

or, in matrix form

$$0 = \theta^T \log \mu = \theta^T u$$

where the logarithms are calculated elementwise

- **Again, the same model structure!**

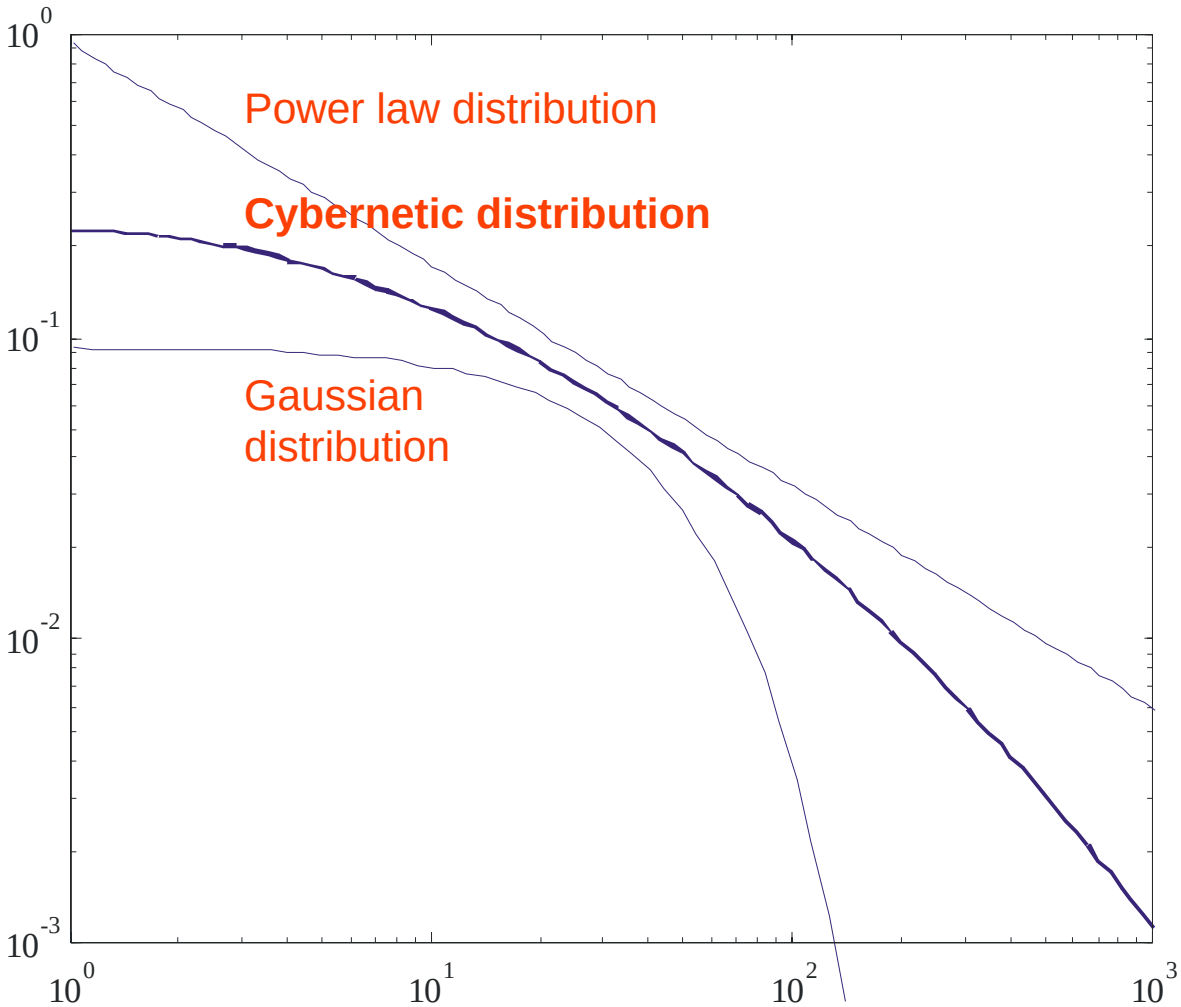
Step aside

- Logarithm of a quantity is a sum of many other logarithms ...
- Assume the numbers being summed are probabilistic
- If they have the same distribution, the central limit theorem applies: Their sum has approximately normal distribution

$$p(\sum_i \log y_i) = c' \exp\left(-\left(\sum_j \log \mu_j - \eta \right)^2 / 2\sigma^2 \right)$$

- The sum has log-normal distribution: On the log/log scale, the distribution of a ”multivariate fractal” quantity behaves quadratically rather than linearly!

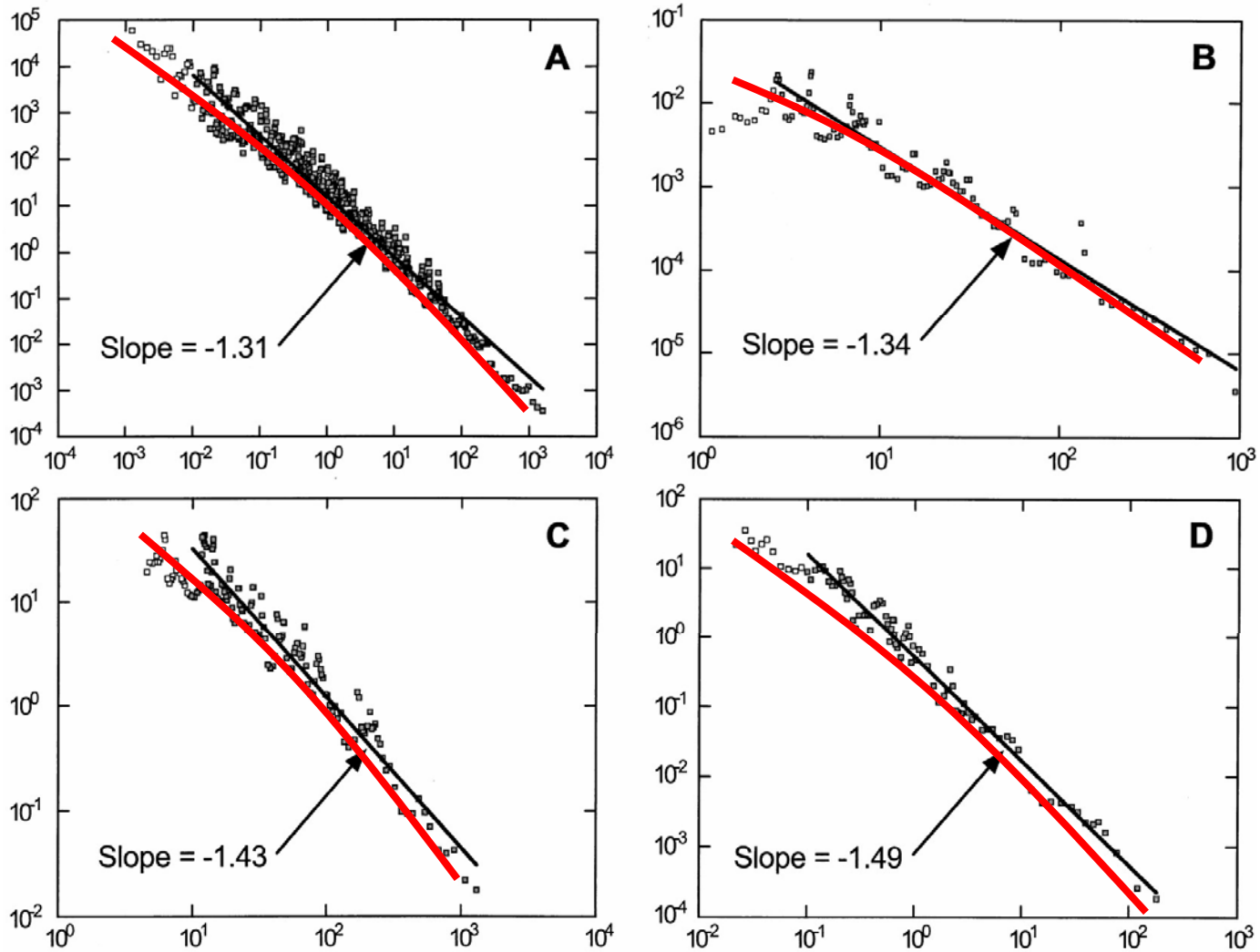
$$\log\left(p(\sum_i \log y_i) \right) = c - \left(\sum_j \log \mu_j - \eta \right)^2 / 2\sigma^2$$



- Longer ”tails” than in normal distribution
- Ends not *so* emphasized as in power law distribution

”You See What You Expect To See”

- Quadratic curves **are** better than linear!



Extents of forest fires (from Science)

- Logarithms are badly behaving – what kind of data vector augmentation to apply instead?
- A simple approximate extension is to include *quadratic* terms among data
- Motivation: First step towards approximating arbitrary nonlinearity (Taylor expansion)
- This choice also supports the notion that multivariate fractal distributions are quadratic in the log/log scale; what is more, the quadratic neocybernetic cost criteria can be globally matched against such variables
- Centering / scaling needed in practice?

$$u' = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \\ \hline \mu_1^2 \\ \vdots \\ \mu_n^2 \\ \hline 1 \end{pmatrix}$$

Models of today's systems: *Constraints*

- As was shown, gene expression networks and metabolic reactions can be based on models of *constraints*

$$0 = \theta^T u$$

  Here, θ is the vector of parameters; when characterizing cell systems, dimension is huge, meaning plenty of parameters
- It is assumed that the data are somehow bound together, and it is this bond that captures the essence of the system
- However, *this constraints-approach is not the only possible*

- How can a (locally linear) model be described otherwise?
- Next, emergent models are studied

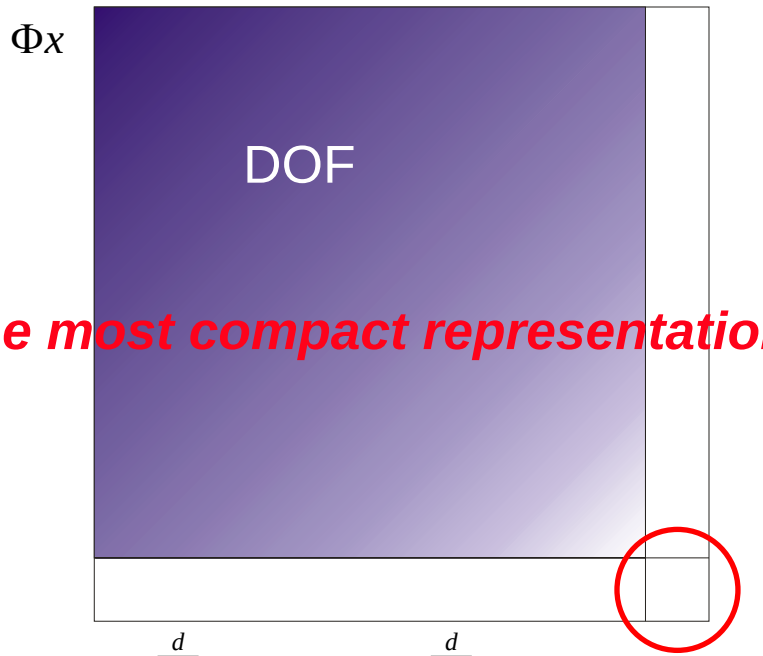
Constraints vs. ”freedoms”

- Claim: The *degrees of freedom* are more characteristic to a system than the constraints are
- Reason: In deeply interconnected systems, emphasis on freedoms is a more compact representation of the system
- The model determines a line in the data space – “null space”, where there is no freedom among data
- “Axes of freedom” = remaining subspace that is *orthogonal to the null space* = basis of a NEW MODEL STRUCTURE
- The eigenvalue decomposition of the data covariance matrix reveals in which directions there is variation in the data and how much: *Eigenvectors = axes of freedom*, and *eigenvalues = their relevances*

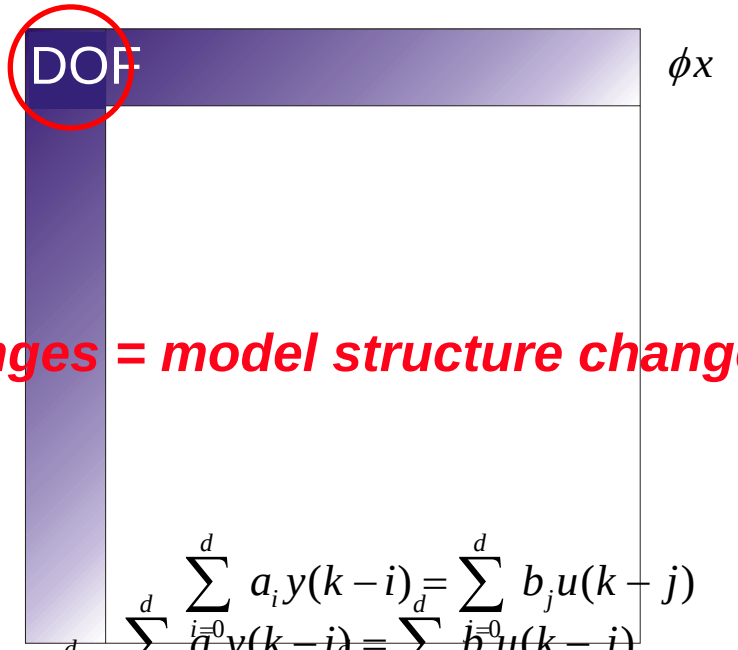
”Emergent Models”

- Data high-dimensional
- Few constraints
- Many degrees of freedom left

- Data equally high-dimensional
- Many constraints
- Few degrees of freedom (right)



$$\sum_{i=0}^{d} a_i y(k-i) = \sum_{j=0}^{d} b_j u(k-j)$$



$$\sum_{i=0}^{d} a_i y(k-i) = \sum_{j=0}^{d} b_j u(k-j)$$

The most compact representation changes = model structure changes

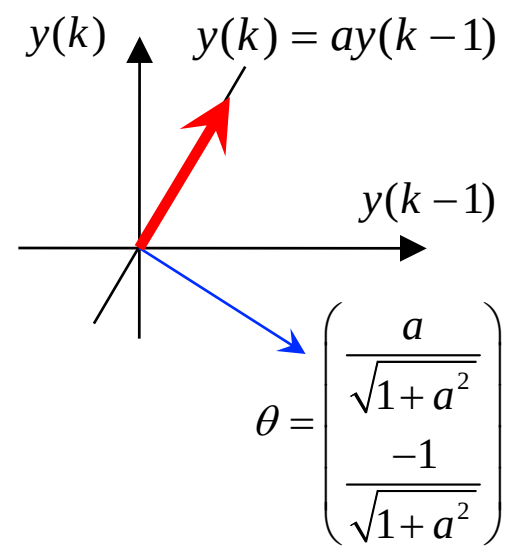
Example

- Assume that

$$y(k) = ay(k-1).$$

Now

$$\theta = \begin{pmatrix} \dfrac{a}{\sqrt{1+a^2}} \\ \dfrac{-1}{\sqrt{1+a^2}} \end{pmatrix}, \qquad u = \begin{pmatrix} y(k-1) \\ y(k) \end{pmatrix},$$



so that

$$S = (\theta \mid \Phi) = \left(\begin{array}{c:c} \dfrac{a}{\sqrt{1+a^2}} & \dfrac{1}{\sqrt{1+a^2}} \\ \dfrac{-1}{\sqrt{1+a^2}} & \dfrac{a}{\sqrt{1+a^2}} \end{array}\right).$$

Normalized basis vectors spanning the whole space $\mathcal{S}$:

Constraint

Axis of freedom

- Now assume (assuming there is redundancy among vars)

$$\begin{cases} y(k) = ay(k-1) \\ y(k+1) = ay(k). \end{cases}$$

In this case (without normalization):

$$\theta' = \begin{pmatrix} a & 0 \\ -1 & a \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad u = \begin{pmatrix} y(k-1) \\ y(k) \\ y(k+1) \end{pmatrix}.$$

The constraint span a two-dimensional subspace in the three-dimensional variable space – one degree of freedom remains

- Orthogonalization of basis θ' (Gramm-Schmidt procedure):

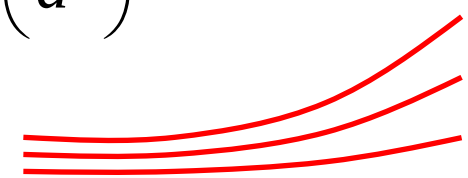
"Deflation"

$$\left(\begin{array}{cc|c} a & 0 & 1 \\ -1 & a & 0 \\ 0 & -1 & 0 \end{array}\right) \rightarrow \left(\begin{array}{c|c} a & \frac{a^2}{1+a^2} & \frac{1}{1+a^2} \\ -1 & \frac{a^3}{1+a^2} & \frac{a}{1+a^2} \\ 0 & -1 & 0 \end{array}\right)$$

(Where is this vector from?)

$$\rightarrow \left(\begin{array}{c|c} a & \frac{a^2}{1+a^2} & \frac{1}{1+a^2+a^4} \\ -1 & \frac{a^3}{1+a^2} & \frac{a}{1+a^2+a^4} \\ 0 & -1 & \frac{a^2}{1+a^2+a^4} \end{array}\right) \longrightarrow \Phi = \begin{pmatrix} 1 \\ a \\ a^2 \end{pmatrix} \Big/ \sqrt{1+a^2+a^4}$$

"Axis of freedom"
exponential form



Prototypes

Towards pattern matching

- Use of the model becomes an *associative pattern matching process* against data (exponential curve in the example)
- Linearity – patterns can be freely scaled and added together
- Vector z is the vector of scaling factors = *latent variables* (note that generally Φ is a *matrix*, containing several ”axes of freedom” as collected together; assume it is *orthonormal*)

$$z(k) = \Phi^T \cdot u(k)$$

- The *reconstruction* where noise is filtered is given as

$$\hat{u}(k) = \Phi \cdot z(k)$$

- The more there are internal constraints (feedbacks, etc.), the more efficient the freedoms-oriented approach becomes

- The pattern matching problem can be expressed also using a cost criterion

$$J = \left(u - \Phi z\right)^T W \left(u - \Phi z\right)$$

- If some of the variables are more relevant than others, the diagonal elements of W are non-identical (for example, one can select $W = \mathrm{Var}\{uu^T\}^{-1}$ or more generally $W = \mathrm{E}\{uu^T\}^{-1}$)
- The solution is

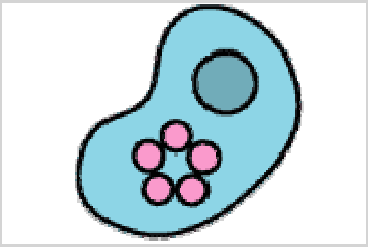
$$\hat{z} = \left(\Phi^T W \Phi\right)^{-1} \Phi^T W \, u$$

- For example, associative regression can be implemented by associatively filling in unknown elements u_i in u: Let $W_{ii} = 0$

”Emergent Models” in practical domains

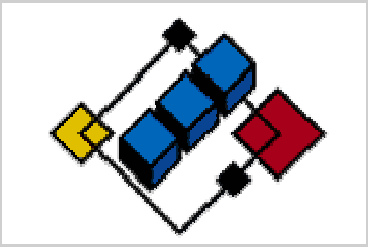
On-going projects:

1. Biological systems: Project ”SyMbolic” in Tekes NeoBio program (here)



   - Plenty of data available, not so much understanding
   - Goal: Systemic models for metabolic and genetic processes
   - Methodology: Extend DOF-thinking to dynamic models

2. Industrial systems: Project ”Testing Manager” in Tekes Äly program



   - Information of components, no understanding of the entity
   - Goal: Models for qualities rather than for individual signals
   - Methods: Extend the DOF-thinking to input/output models

Freedoms: Extension to chemical domain

- It is not only ”programmed” systems where the freedoms-based thinking applies: In chemical equilibrium systems, the linear constraints also apply (as shown before)
- In complex systems, not all reactions are known – this does not matter as long as the system remains stable, one can concentrate on the freedoms
- One ends in familiar methods, seen from another point of view: Le Chatelier principle states that changes in environment are compensated by changes in the balance
- The leap is conceptual: Linear modeling of balances is not only data modeling but system modeling

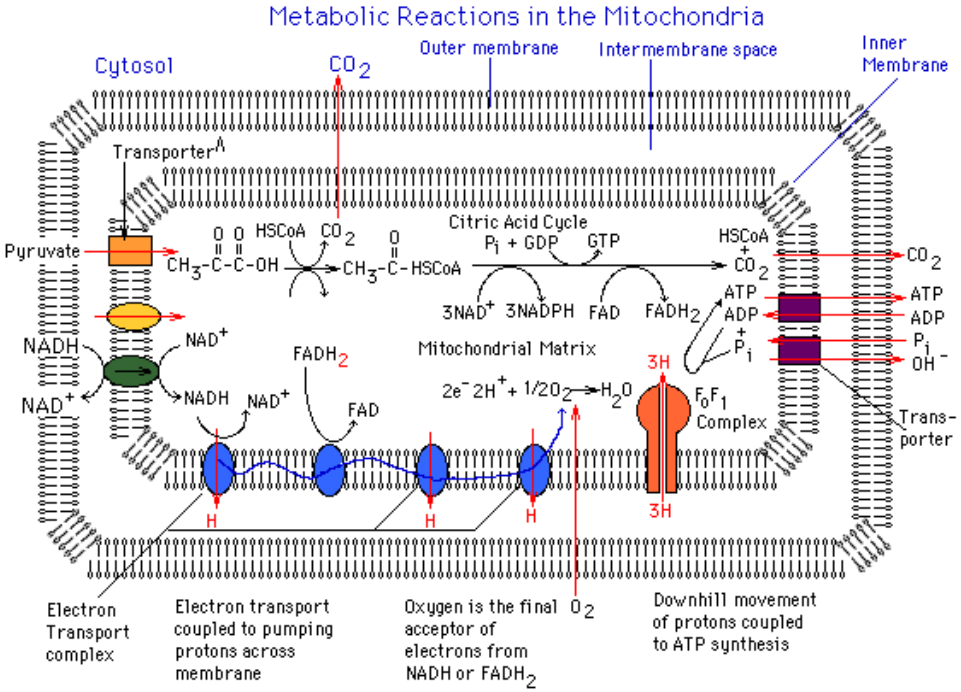
- Extending biological cybernetics to technical bioprocesses: The still unbounded degrees of freedom can be regulated?

”Superorganisms” constructed by added external feedbacks!

Cell level #1: Metabolic system

- Constraints = Balance equations
- DOF's = Metabolic behaviors or functions?
- Anthropocentric interpretations: Nutrient, waste product

- When complexity cumulates, the balance reactions start looking goal-oriented, pre-planned, and "clever"
- For example, scarcity of some chemical changes the balance appropriately

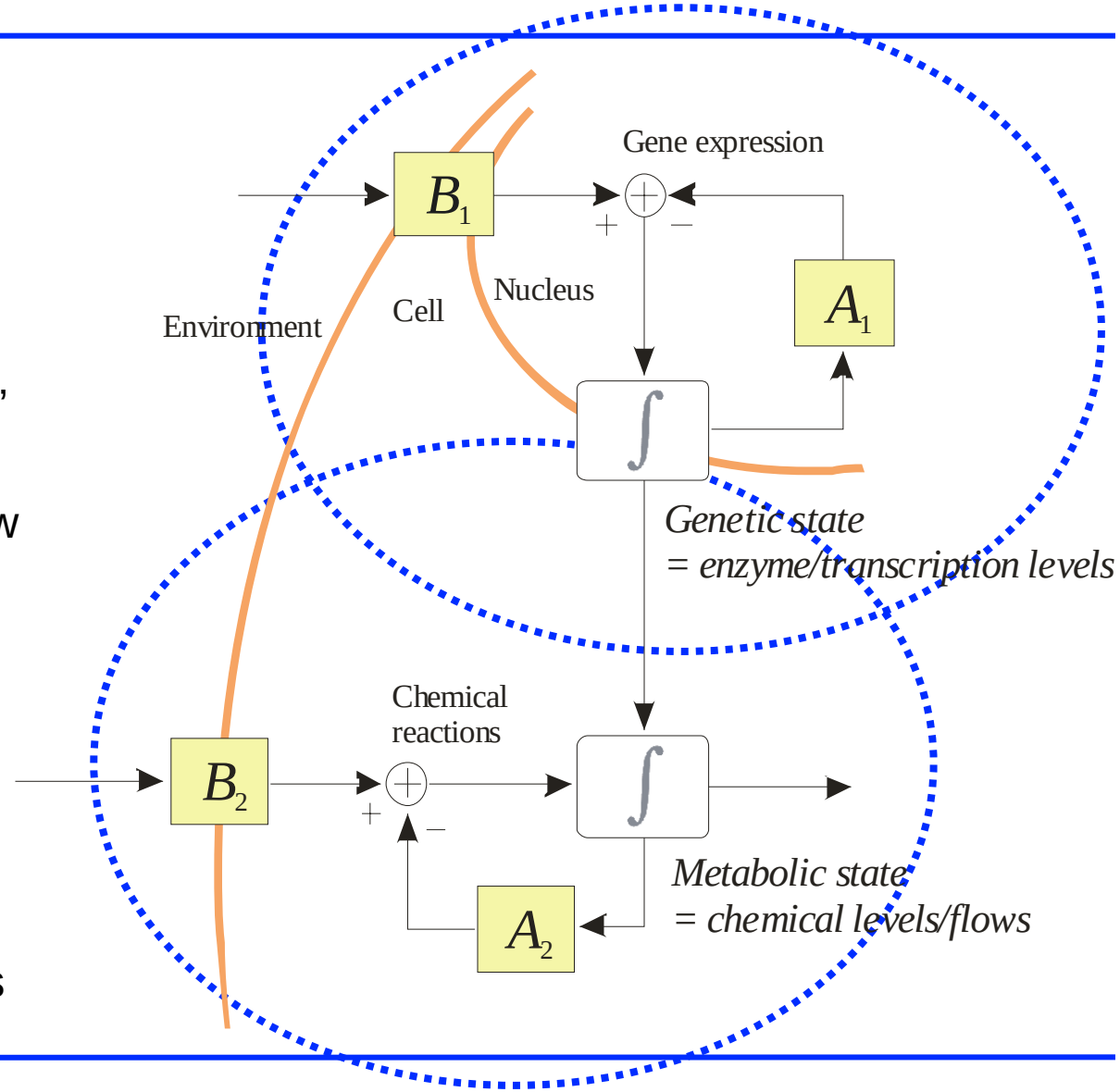


Cell level #2: Genetic system

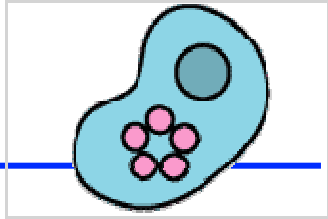
- Active genes determine the enzymes (proteins) available = the reactions actually taking place in the cell
- Special enzymes act as *transcription factors*, activating (or inhibiting) other genes
- The gene activation relationships constitute a causal network – Again: Assume ”pancausality”
- In equilibrium, causal ”forces” balance each other even though the underlying processes are very complex
- The same abstraction applies to all processes where DNA, mRNA, or the enzymes are pre- and postprocessed
- Static model rather than a huge set of sequential, elementary ones

... Two cybernetic levels of cell processes

- Appropriate abstractions:
  - Two successive process levels of ”generalized diffusion”
  - Metabolic processes fast, genetic ones slow
  - In both cases, forget sequential nature
  - Both levels – same approaches – can be combined in one!?
  - Emergent models based on latent (logarithmic) variables



Project ”SyMbolic”



Systemic Models

for Metabolic Dynamics and Gene Expression

- Co-operating partners: Helsinki University of Technology (Control Engineering Laboratory, and Neural Networks Research Centre), Finnish IT Center for Science CSC, and MediCel Oy
- Time span 2004 – 2005+
- Funding by Tekes NeoBio Program

- *Here, the starting points applied at HUT Control Engineering Laboratory are presented + preliminary results*

Assumptions now

- Study living cell rather than pathological (irrelevant?) cases (no kick-off experiments?)
- Balances are more characteristic than transients; steady states are (first) modeled
- Metabolic processes are well buffered: Linear models are locally applicable
- Rather than carrying out tests in a SISO manner, the whole grid of proteomics/metabolomics are studied simultaneously
- Around the operating point gene expression is a part of the metabolic system: Gene activities included in the data vector
- In principle, the same preprocessing of data applicable on both level data: Logarithms or relative changes

What is being done

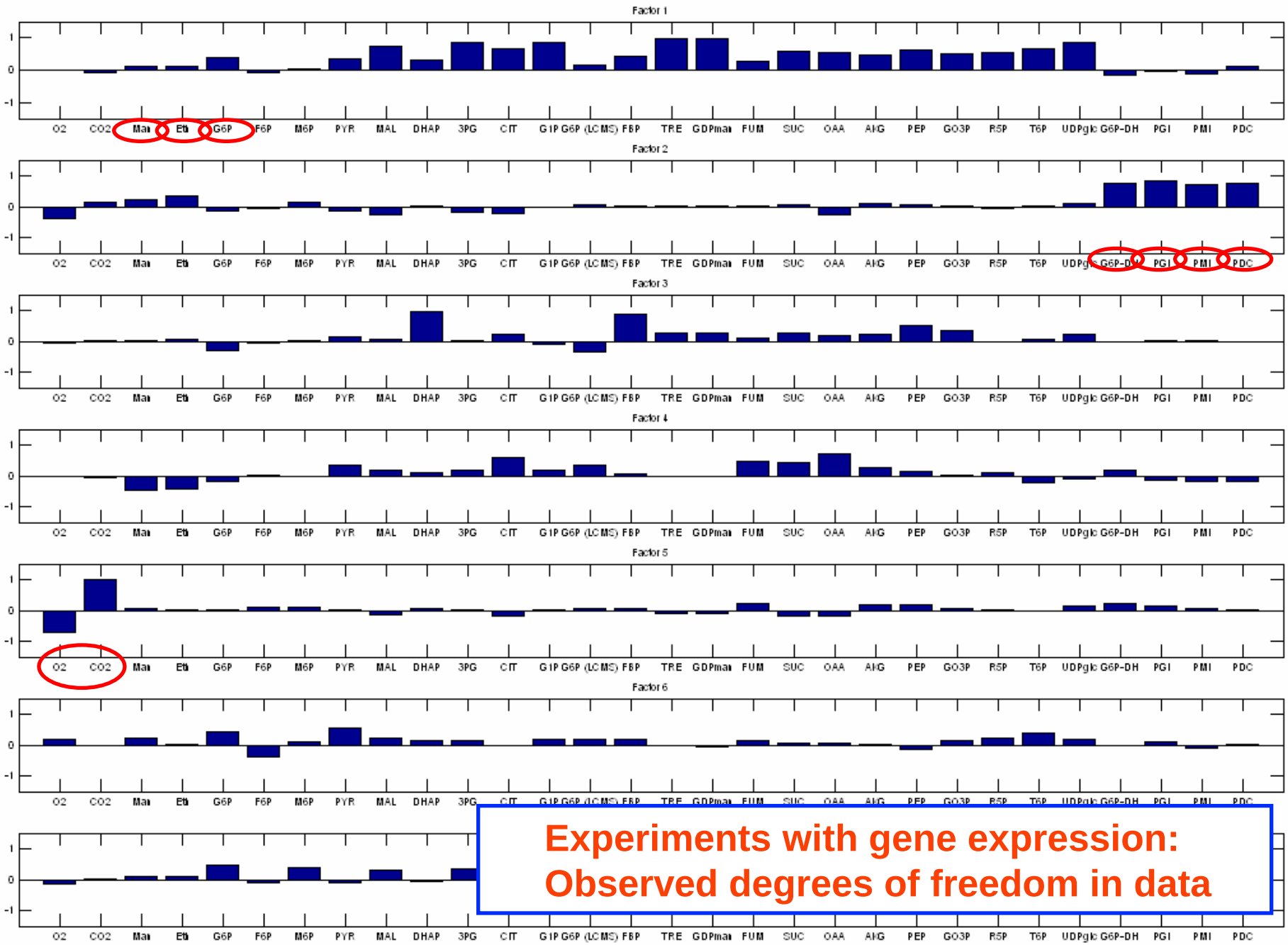
- Define the sets of metabolites, transcription factors, and relevant environmental conditions (temperature, pH, ...)
- Carry out experiments in different conditions, collect data during the transient and in steady state

- Find the degrees of freedom, determining the metabolic functions and genetic functions
- Match the transient data to determine *dynamic* parameters: *Subspace identification*, or *Maximum Entropy Pursuit* principle applied

- Later – extending the analyses, proceed towards nonlinear, *sparse-coded* models?

Challenges

- Data collection is typically carried out applying SISO-type tests:
  - Because of buffering (balance pursuit) huge dosages are needed in the single input – minor effects in one, and all change – or cell crippled
  - Single ”kick-off genes” are explicitly deactivated, resulting in non-natural behaviors
- Practical problems:
  - Metabolics are difficult to measure (however, gene activities found applying ChIP technique)
  - Different formalisms, incompatible tests in data warehouses
- Additionally: ”Extra” behaviors become visible in stress (transient) situations



Experiments with gene expression: Observed degrees of freedom in data

Dynamic *open-loop* experiments

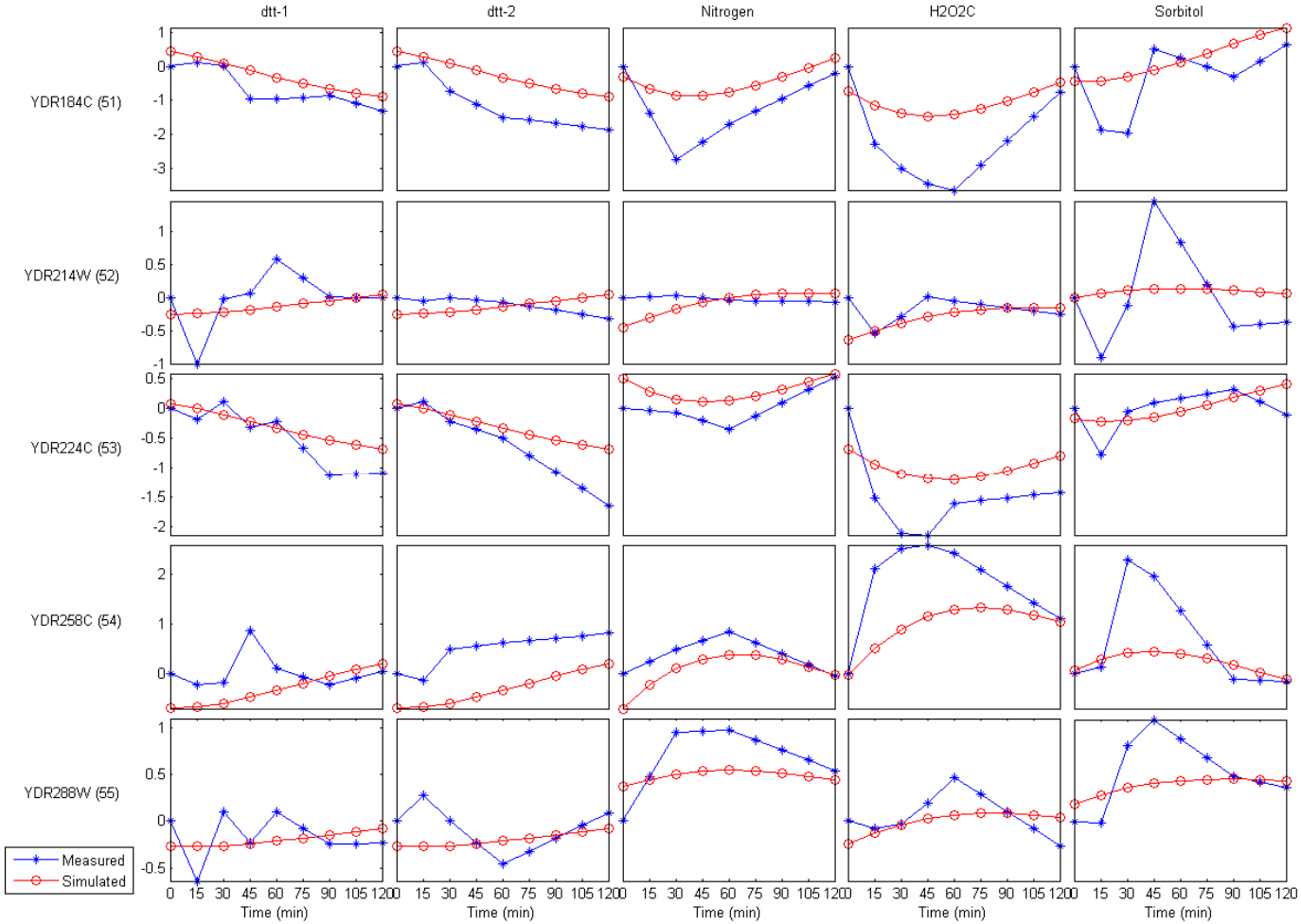
- ***n* = 4** only!
- dim(u) = 10 (+ 10 nonl.)
- dim(y) = **4135** (all!)

Data from:

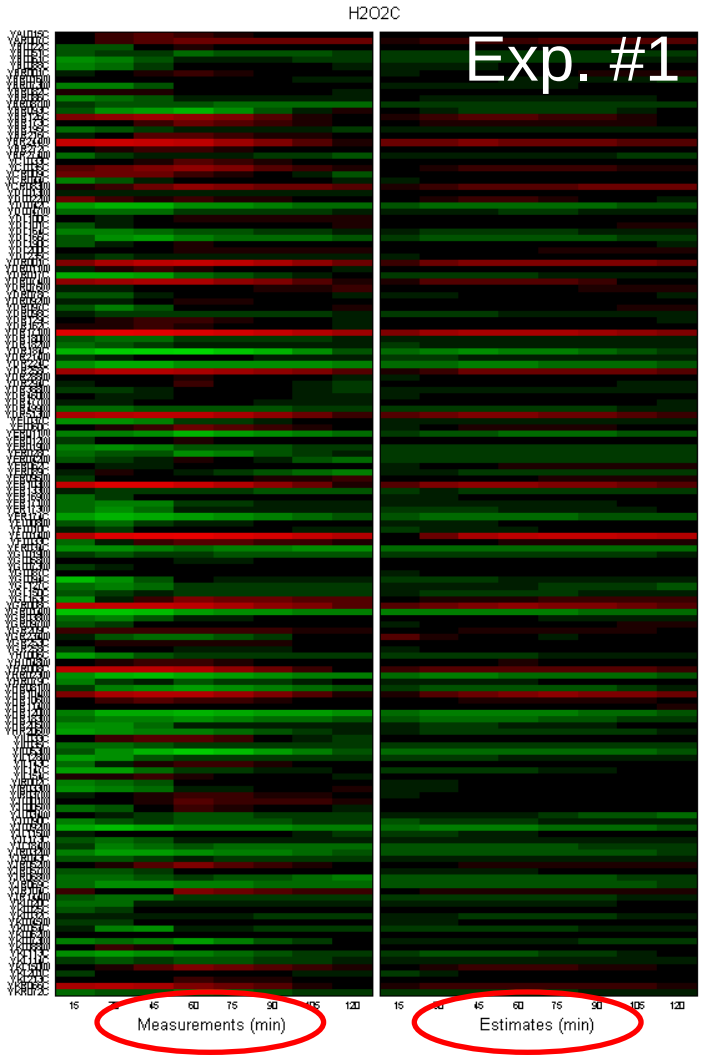
A. P. Gasch *et al*: Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, Molecular Biology of the Cell, Vol. 11, 4241-4257, December 2000

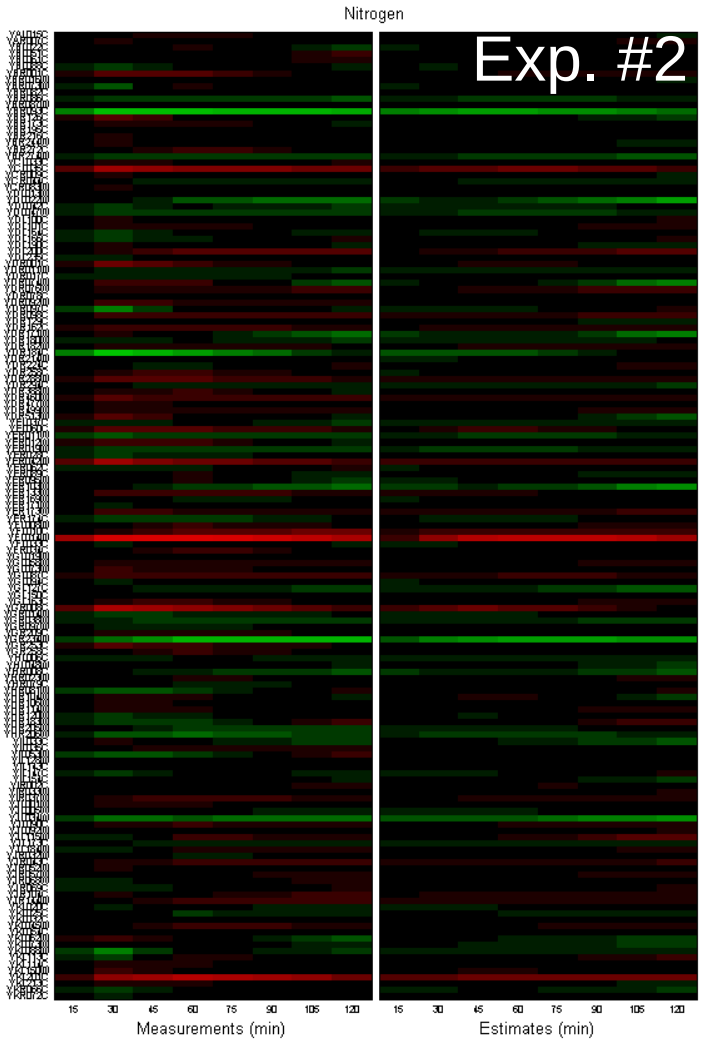
and

H. C. Causton *et al*: Remodeling of Yeast Genome Expression in Response to Environmental Changes, Molecular Biology of the Cell, Vol. 12, 323-337, February 2001



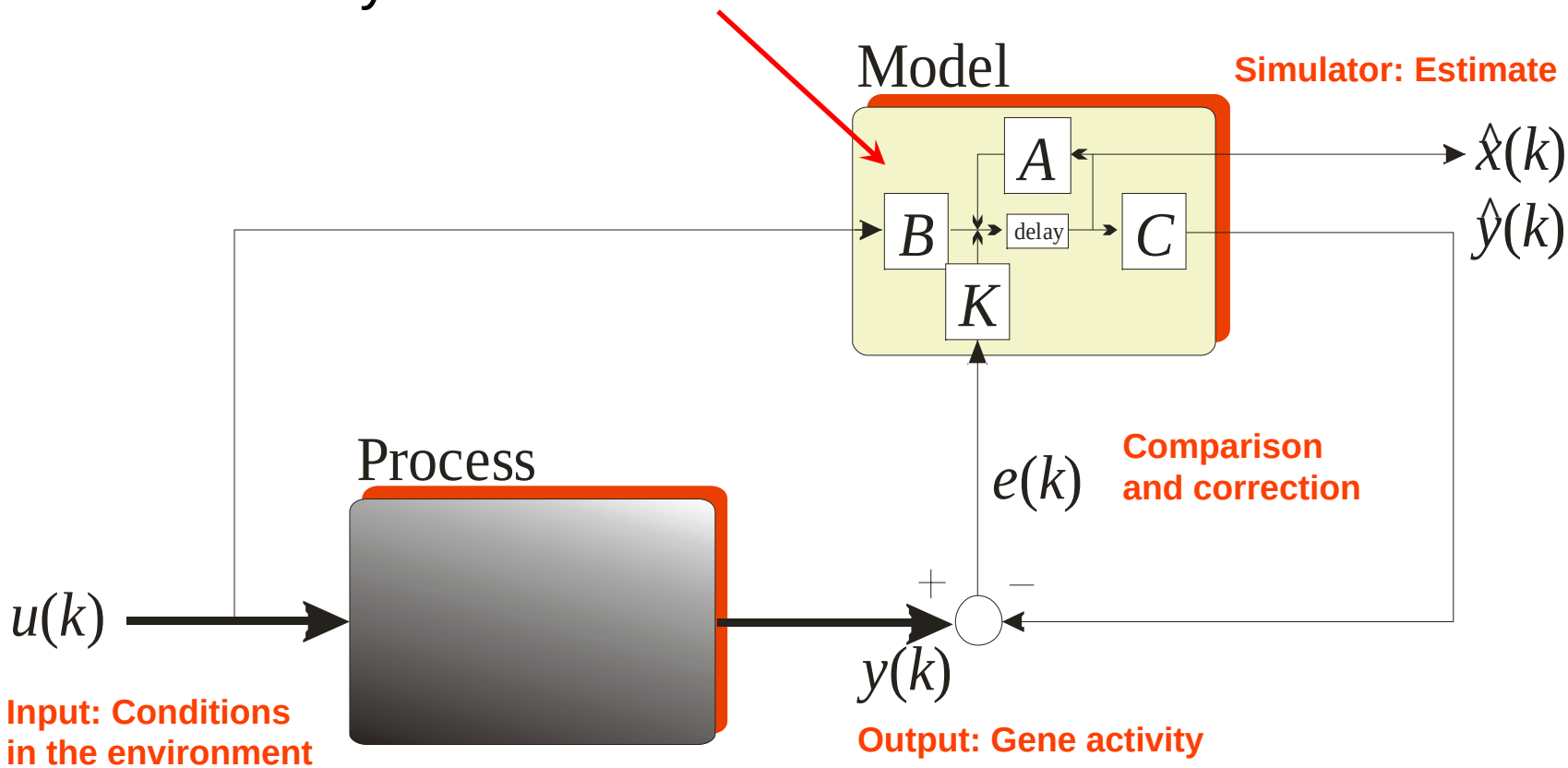
- 254 ”stress genes” shown





Practical estimation of state and output

- Discrete model construction applying *subspace identification*
- Estimation by *Kalman filter*



Conclusion

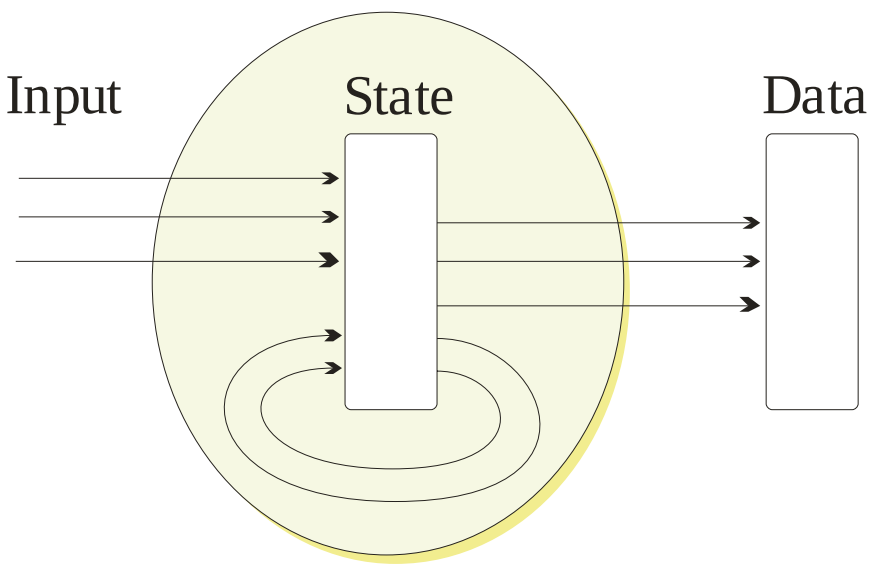
- Cybernetic models have the same structure, no matter if they are based on networks or explicit constraints:

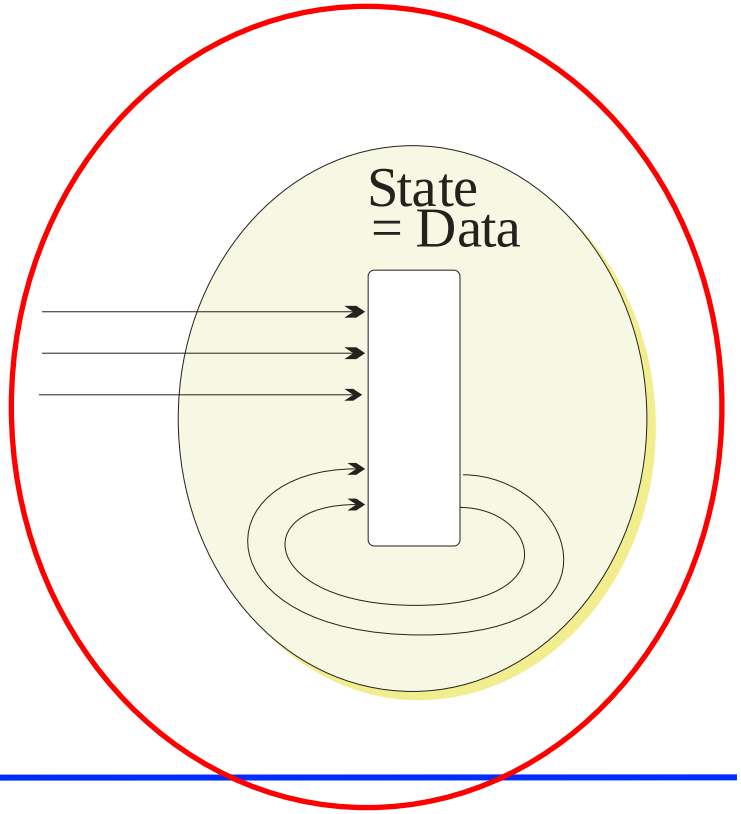
  Locally linear reduced-dimension latent variable subspace, multivariate models representing dynamic equilibria

- All systems can be studied in the same framework – applying PCA / factor analysis, etc.
- Data preprocessing = augmenting the data space
- High dimensionality, redundancies, noise, etc., are efficiently tackled with
- The model is not only a data model, it is a **system model**, meaning that its predictions can be applied for design?!

”System model” vs. ”data model”

- Observed dependencies among variables are not only *correlations* but *causalities*





Extensions ...

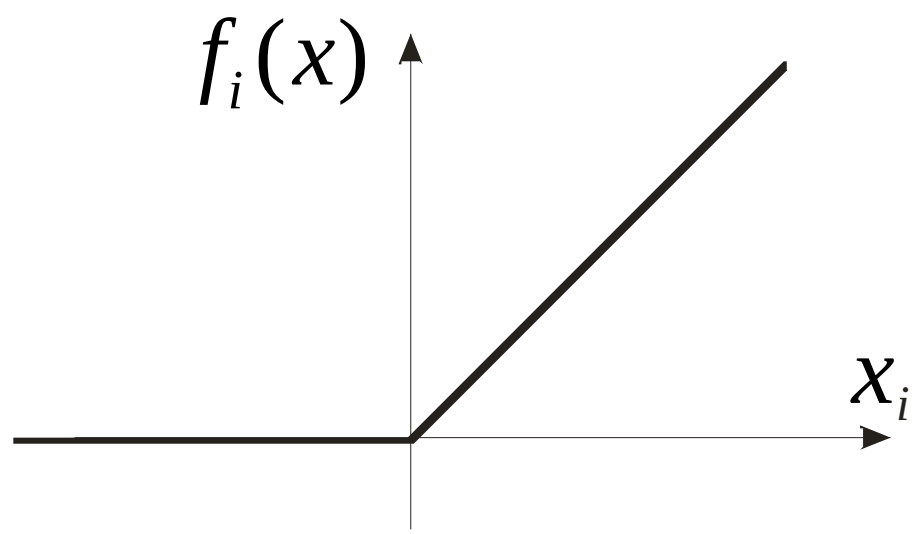
- Genetic system implements the reservoir of prototypical structures to choose among
- Genes are either inactive, or they are active in varying degrees, determining which structural alternatives are employed, and to what degree
- Active gene combinations determine the ”basins of attraction”, and continuous state optimization within these structural constraints takes place
- Genes seem to be highly redundant, making the inherently discontinuous coding look more continuous
- Gene activity/inactivity can be modeled in the same way as chunk activity/inactivity in the cognitive system?

How to implement ”state-controlled” sparsity in a mathematical model?

CUT function

- Linearity being the starting point – gives intuitions in which directions to extend the framework
- A simple example of nonlinear extensions: CUT function
- If variable is positive, let it through; otherwise, filter it out

$$f_i(x) = \begin{cases} x_i, & \text{kun } x_i > 0 \\ 0, & \text{kun } x_i \leq 0. \end{cases}$$



Negative signals are not "visible"

Simulations: “Winner-Take-All”

- Simulation model

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} -\gamma & -1 \\ -1 & -\gamma \end{pmatrix} \cdot \begin{pmatrix} x_{\text{cut},1}(t) \\ x_{\text{cut},2}(t) \end{pmatrix} + \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix} \cdot \begin{pmatrix} x_{\text{in},1}(t) \\ x_{\text{in},2}(t) \end{pmatrix}$$

- $C_1/C_2 = 0.95$, and vice versa