

## **Challenges of Complexity and Integration in Quantitative Systems Biotechnology.**

Eberhard O. Voit  
Department of Biometry and Epidemiology  
Medical University of South Carolina  
[VoitEO@MUSC.edu](mailto:VoitEO@MUSC.edu)

It is widely acknowledged that the enormous amount of data produced by the human and other genome projects requires new mathematical and computational strategies of analysis and integration. Classified very broadly, these strategies may either address large-scale databases and networks, or they may focus on the intricate details of smaller regulatory systems. Eventually, the two approaches must converge, but our current methodologies may not be sufficient for this to happen now. At this point, typical large-scale approaches are designed to detect relationships among genes or between gene expression and function. These approaches make extensive use of the ability of computers to address combinatorial problems with high efficiency. Although very successful in many respects, these approaches alone are not sufficient. They must be complemented with detailed algebraic and numerical analyses that aim at discovering and explaining the design principles behind natural systems and at integrating diverse pieces of information within and between levels of biological organization. It has been shown that such analyses can help explain why a gene circuit or a metabolic pathway is regulated in particular way and not in another, theoretically possible fashion. Detailed smaller-scale analyses can lead to a rationale for why genes of the same pathway may be over-expressed at drastically different rates, when the organism is exposed to a stimulus. They may provide reasons for why the artificial over-expression of genes in a biotechnological setting does not necessarily result in the desired and expected increase in product yield. The presentation primarily discusses the challenges of integration in complex systems and briefly mentions a mathematical approach, based on power-law approximation, that has been helpful in dealing with some of these challenges.

## **Challenges of Complexity and Integration in Quantitative Systems Biotechnology**

Eberhard O. Voit  
Department of Biometry and Epidemiology  
Medical University of South Carolina  
1148 Rutledge Tower  
Charleston, SC 29425  
VoitEO@MUSC.edu

National Science Foundation  
13-14 September 2000

### **Questions of Interest**

- What are the unmet (near term) opportunities and needs associated with using genomic information for phenotype prediction?
- What are the developments needed in test-bed and high throughput screening systems?
- What new mathematical tools, or those that can be adopted from other disciplines, are needed to use genomic information for predictive purposes?

## Opinion Statement

Unmet (near-term) opportunities and needs associated with using genomic information arise from the **complexity of organisms**.

Mathematical tools must **aid our understanding of complexity**. They must facilitate the **integration** of diverse types of information in conceptual and **quantitative frameworks**.

## Characteristics of Complexity

Large numbers of components

Large number of processes

Processes are nonlinear

Quantitative changes in parameters cause qualitative changes in response

## **Large Numbers of Components and Processes**

Biologists have accumulated vast information

Computer scientists have developed management tools

Bioinformatics promises huge increases in new data

- ⇒ **Data acquisition seems “under control”**
- ⇒ **Should be funded from other sources**

## **Nonlinear Processes**

Simple extrapolation often faulty

Cause-and-effect thinking insufficient

Superposition not necessarily valid

- ⇒ **Needs effort**

## Quantitative Changes in Parameters

$p < \vartheta \Rightarrow$  system moves to steady state

$p \geq \vartheta \Rightarrow$  system oscillates

System responses difficult to predict

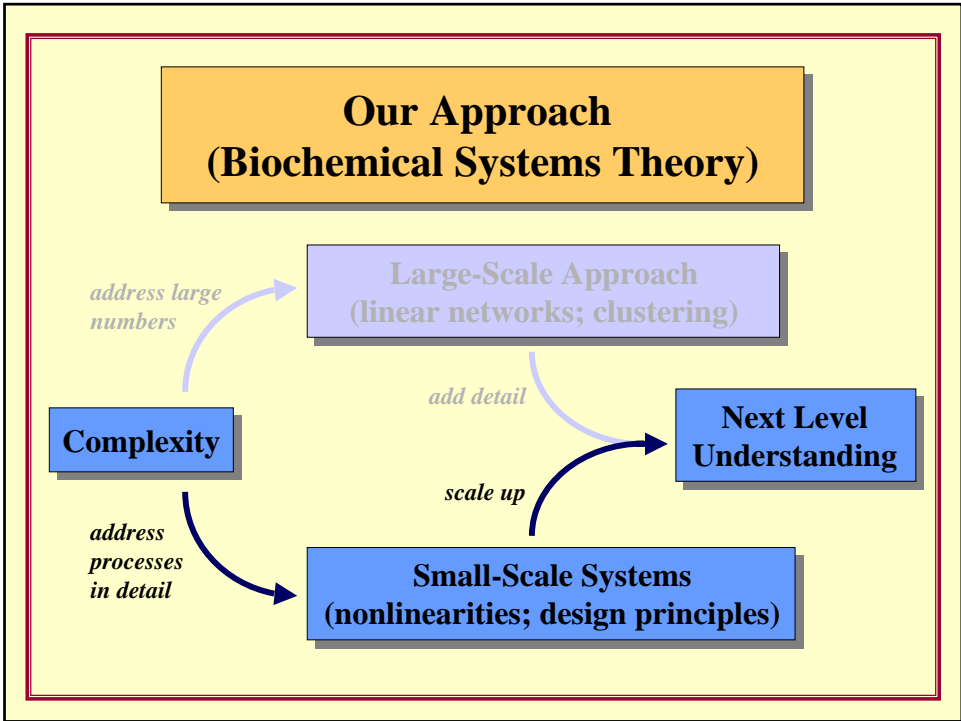
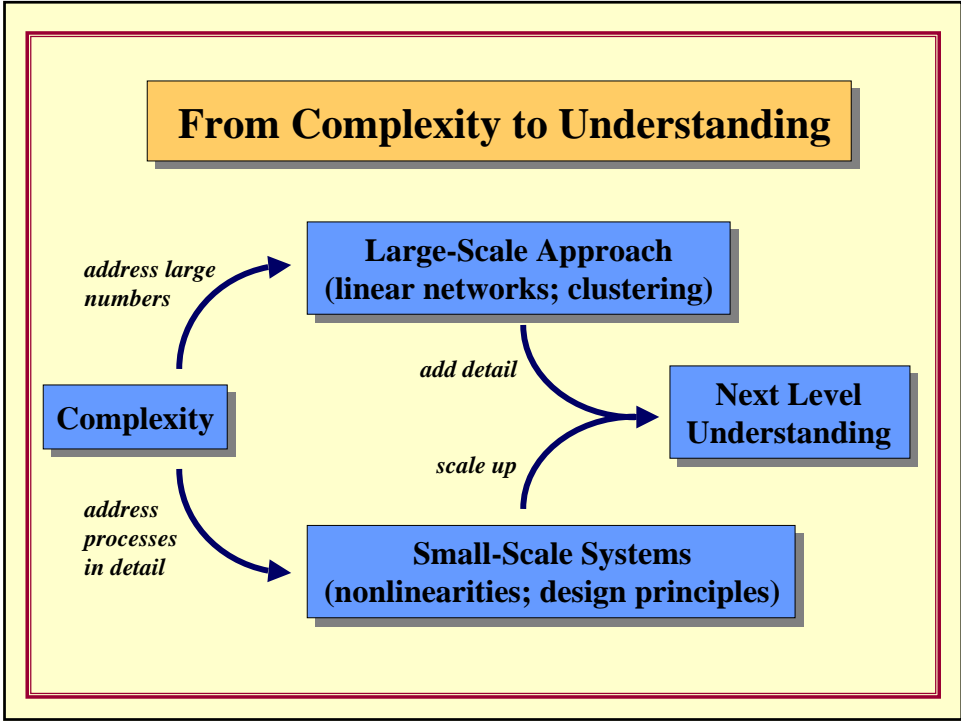
**$\Rightarrow$  Needs effort**

## From Complexity to Understanding

Strategy 1: Immediately address systems of large size;  
minimize need for (and consideration of)  
detail

Strategy 2: Develop true understanding of small systems;  
scale up

**$\Rightarrow$  Strategies 1 and 2 should both be pursued;  
they should complement each other.**



## Large-Scale Approaches

Multitude of algorithms for sequencing, gene finding, clustering; multitude of databanks

⇒ **Need better statistics, noise reduction**

⇒ **Leave algorithms and databanks to industry?**

Large models: stoichiometric networks, *E-cell*, *Entelos*

⇒ **Need effort**

## Limitations of Large-Scale Approaches

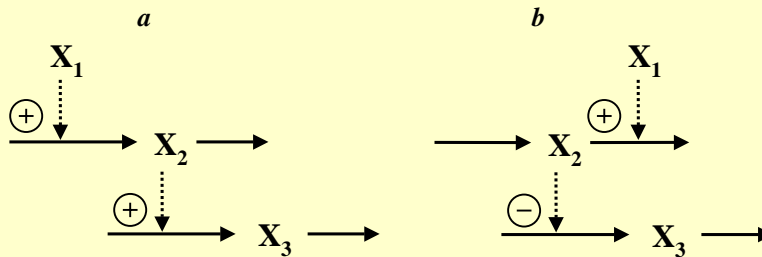
*Fail to provide explanations of structural details:*

Example: Alternative designs

Example: Over-expression of genes

Example: Yield optimization in biotechnology

## Alternative Designs (Savageau)



Two different regulatory control structures show outwardly equivalent responses to changes in  $X_1$ . Why didn't nature eliminate "redundant" designs?

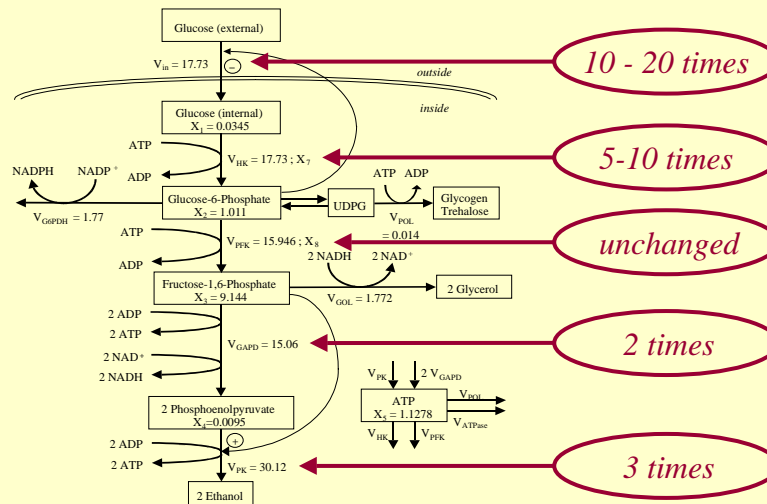
## Over-Expression of Genes

Why are genes that code for enzymes of the same pathway over-expressed at significantly different rates?

Clustering cannot explain this observation.

EOV and Radivoyevitch (2000) provided explanations based on a detailed biochemical model.

## Example: Glycolysis in Heat-Shocked Yeast



## Yield Optimization in Biotechnology

**Goal: Over-express genes/enzymes for increased yield**

Ruijter *et al.* failed with intuitively reasonable, biotechnologically well-executed approach.

Torres *et al.* (1996-2000) used *Biochemical Systems Theory* to prescribe optimal over-expression patterns.

They showed that alterations in only one, two or three control variables are ineffective.

EOV and Del Signore showed that imprecise over-expression of enzymes does not improve yield.

## Small-Scale Systems

### Goals:

Understand design principles

Facilitate up-scaling to large systems

Understand integration within and between levels of organization

⇒ **Discuss Biochemical Systems Theory (BST) in the following**

## Biochemical Systems Theory

Ordinary differential equations; one for each dependent variable; represent influxes and effluxes as products of power-law functions:

$$dX_i/dt = V_i^+(X_1, \dots, X_n) - V_i^-(X_1, \dots, X_n)$$

becomes *S*-system:

$$dX_i/dt = \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \dots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \dots X_n^{h_{in}}$$

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \quad i = 1, 2, \dots, n$$

## Advantages of *BST* Representation

Very general: Allows for essentially any (smooth) nonlinearity, including stable oscillations, chaos.

Steady-state equations linear.

Structure permits powerful diagnostics.

Structure permits very efficient numerical analysis (*PLAS*).

Structure is readily scaled up to arbitrary size.

## Steady-State Equations

Steady state equation:

$$\alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} = \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \quad i = 1, 2, \dots, n$$

Define  $y_i = \ln X_i$ ,  $b_i = \ln(\beta_i / \alpha_i)$ ,  $a_{ij} = g_{ij} - h_{ij}$ . Steady-state equations become

$$a_{i1}y_1 + a_{i2}y_2 + \dots + a_{in}y_n + a_{i,n+1}y_{n+1} + \dots + a_{i,n+m}y_{n+m} = b_i \quad i = 1, 2, \dots, n$$

$$\mathbf{A}_D \cdot \mathbf{y}_D + \mathbf{A}_I \cdot \mathbf{y}_I = \mathbf{b}$$

## Evaluation of Steady-State Equations

Solution (if steady-state point exists):

$$\mathbf{y}_D = \mathbf{A}_D^{-1} \cdot \mathbf{b} - \mathbf{A}_D^{-1} \cdot \mathbf{A}_I \cdot \mathbf{y}_I$$

Quantities in the equation can be read off directly from system equations.

All properties of the system close to st.st. are consequences of this equation.

Uniqueness depends on rank of  $\mathbf{A}_D$ .

Stability characterized by eigenvalues.

## System Diagnostics

System responses and sensitivities derived from steady-state equation

$$\mathbf{y}_D = \mathbf{A}_D^{-1} \cdot \mathbf{b} - \mathbf{A}_D^{-1} \cdot \mathbf{A}_I \cdot \mathbf{y}_I$$

For instance, change in rate constant changes  $\mathbf{b}$ ; strength of effect on  $y_D$

$$\frac{\partial \mathbf{y}_D}{\partial \mathbf{b}} = \mathbf{A}_D^{-1} \quad (\text{parameter sensitivities})$$

Change in an independent variable amplified/attenuated as

$$\frac{\partial \mathbf{y}_D}{\partial \mathbf{y}_I} = -\mathbf{A}_D^{-1} \cdot \mathbf{A}_I \quad (\text{signal propagation; gains})$$

Algebraic analysis possible (in principle); numerical analysis very efficient.

## System Diagnostics (cont'd)

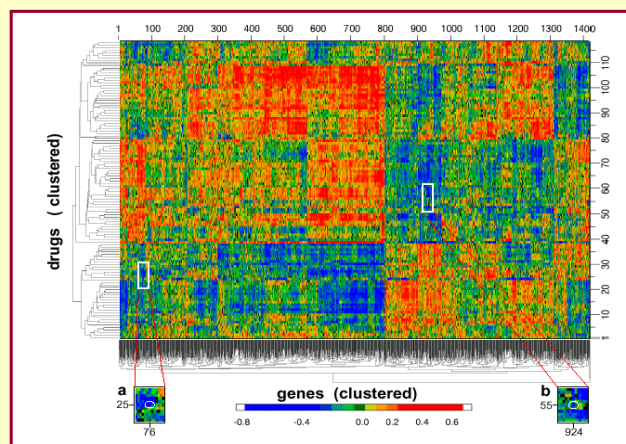
Large sensitivities (and gains) are often signs of problems:  
System is not robust.

Location of large sensitivities (and gains)  
pinpoint problematic components or subsystems.

Iteration between diagnostics and model refinement  
leads to better (“optimal”?) model.

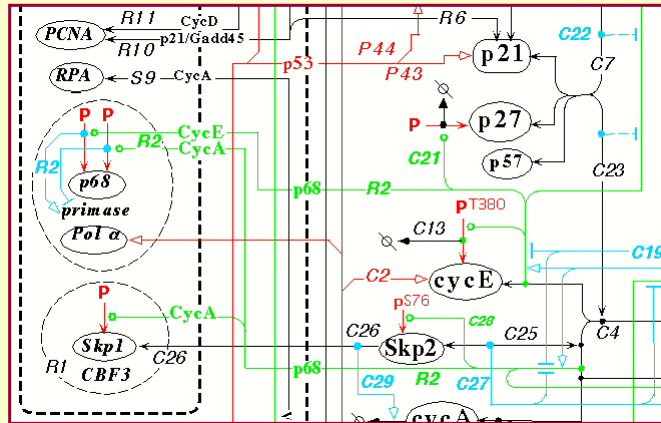
Example: Sequence of models for purine metabolism  
(Curto *et al.*, 1997, 1998ab; Voit, 2000)

## Integration Within Levels



Why are particular genes over-expressed simultaneously?

## Integration Within Levels



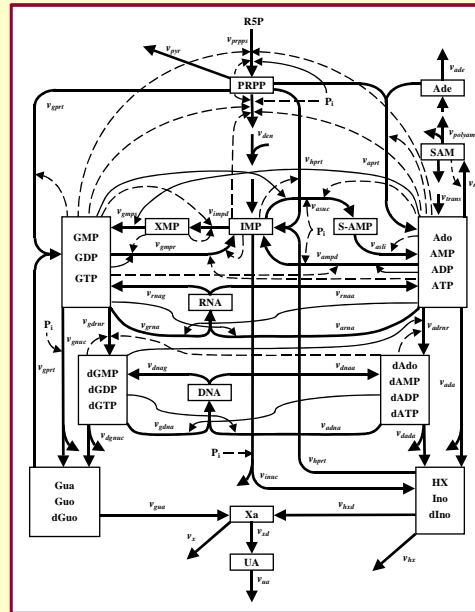
How is gene expression regulated and coordinated?

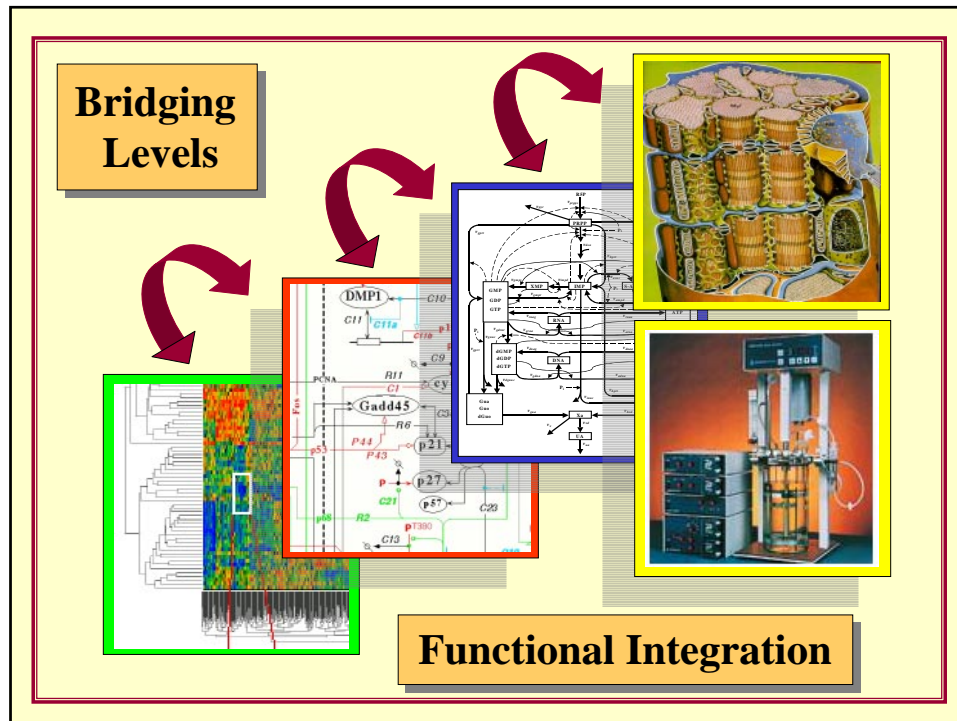
## Integration Within Levels

How are biochemical pathways regulated and coordinated?

Why are biochemical systems designed in a particular fashion?

What do we need to know to design biochemical pathways from scratch?





## Summary

Unmet (near-term) opportunities and needs associated with using genomic information arise from the **complexity of organisms**.

Mathematical tools must **aid our understanding of complexity**. They must facilitate the **integration** of diverse types of information in conceptual and **quantitative frameworks**.