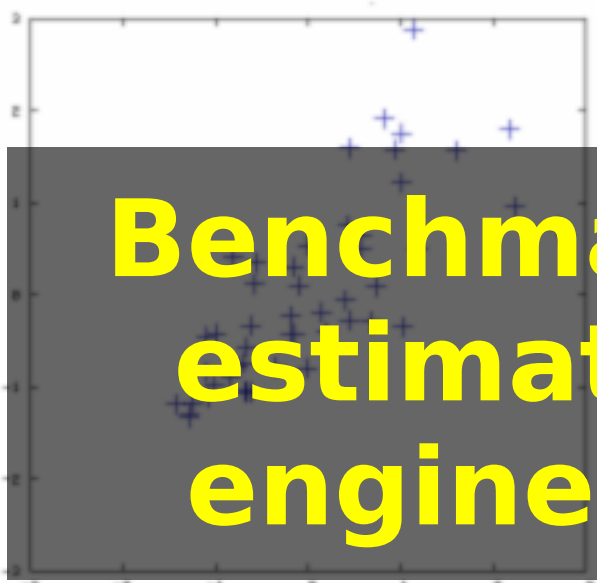
Benchmarking parameter estimation and reverse engineering strategies

Pedro Mendes, Diogo Camacho, Paola Vera-Licona, Reinhard Laubenbacher

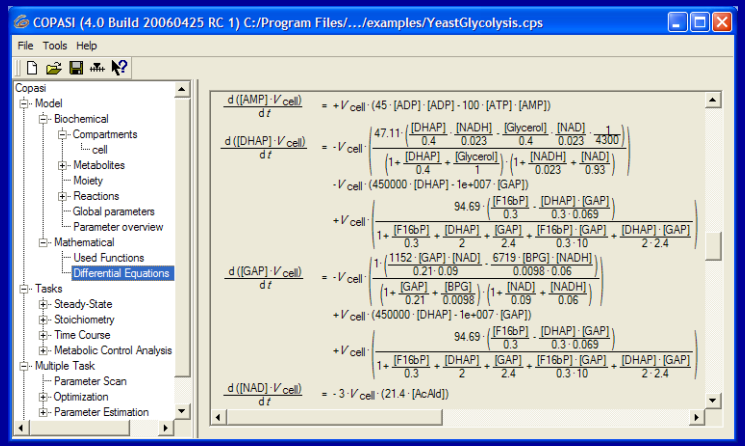
http://mendes.vbi.vt.edu

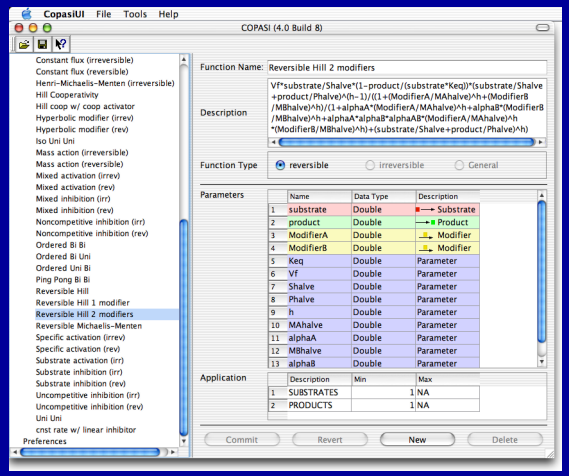


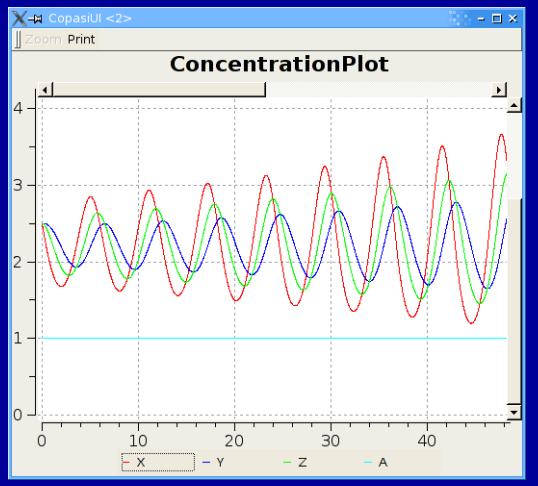
copasi

- ODE and stochastic simulation
- MCA, stability analysis, parameter scans
- Optimisation, fitting
- Sensitivity analysis
- GUI and command line versions
- Reads/writes SBML

- Mendes group
- Kummer group







http://www.copasi.org

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 24 2006, pages 3067–3074
doi:10.1093/bioinformatics/btl485

Systems biology

COPASI—a COmplex PAthway SImulator

Stefan Hoops[1,†], Sven Sahle[2,†], Ralph Gauges[2], Christine Lee[1], Jürgen Pahle[2], Natalia Simus[2], Mudita Singhal[1], Liang Xu[1], Pedro Mendes[1,*] and Ursula Kummer[2]

[1]Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, VA 24061, USA and [2]Bioinformatics and Computational Biochemistry, EML Research, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany

Received on June 29, 2006; revised on August 29, 2006; accepted on September 14, 2006
Advance Access publication October 10, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Simulation and modeling is becoming a standard approach to understand complex biochemical processes. Therefore, there is a big need for software tools that allow access to diverse simulation and modeling methods as well as support for the usage of these methods.
Results: Here, we present COPASI, a platform-independent and user-friendly biochemical simulator that offers several unique features. We discuss numerical issues with these features; in particular, the criteria to switch between stochastic and deterministic simulation methods, hybrid deterministic–stochastic methods, and the importance of random num-

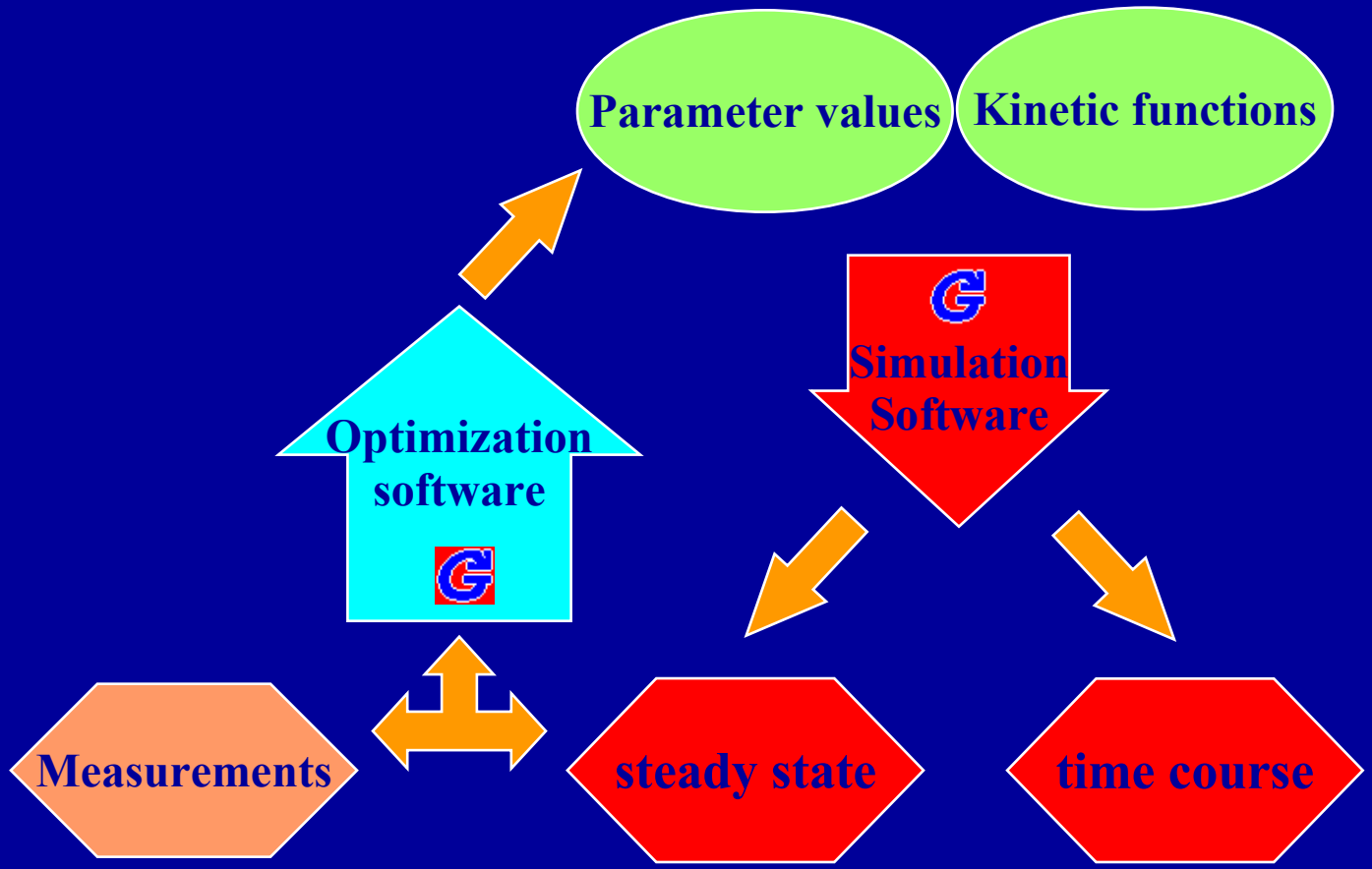
and flux analysis (Klamt *et al.*, 2003). However, some tools contain whole suites of functionalities, e.g. simulation, flux and control analysis (Tomita *et al.*, 1999; Sauro *et al.*, 2003; Meng *et al.*, 2004).

In order to improve the compatibility of these tools, markup languages such as SBML (Hucka *et al.*, 2003) and CellML (Lloyd *et al.*, 2004) were created to allow model exchange. Many tools are now able to read and write models in these file formats.

Here we present a new program—COPASI (COmplex PAthway SImulator)—which combines all of the above standards and some unique methods for the simulation and analysis of biochemical

Hoops *et al.* (2006) *Bioinformatics* **22**, 3067-3074

Parameter estimation



Optimization methods

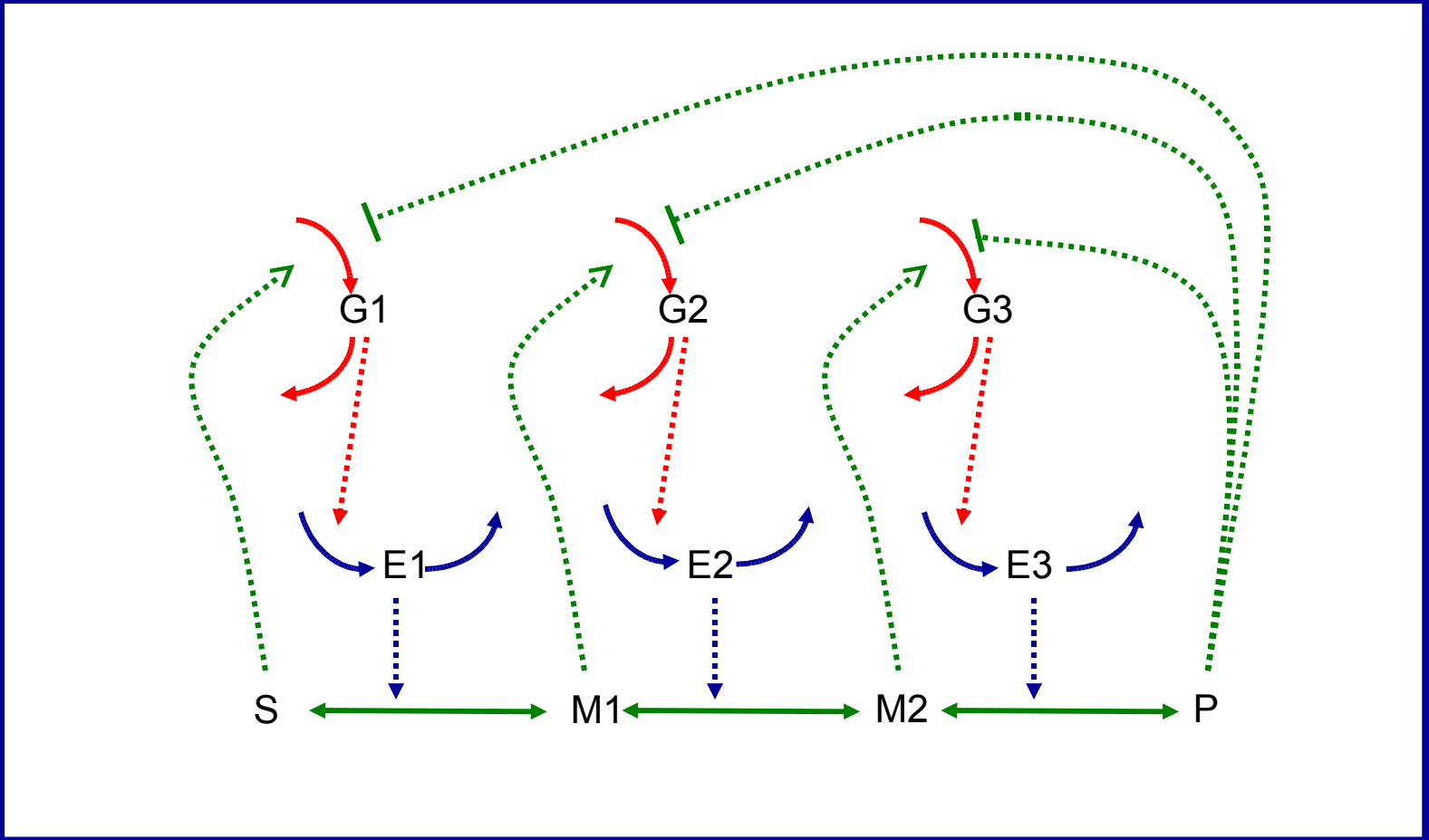
Based on derivatives:

- L-BFGS-B
- Levenberg-Marquardt
- Steepest descent
- Truncated Newton
- NL2SOL
- Derivative Tensor method

Direct search:

- Hooke & Jeeves
- Nelder & Mead (simplex)
- Multistart
- Genetic algorithm
- Evolutionary programming
- Simulated annealing
- Particle swarm
- Random search

3-step pathway with gene expression



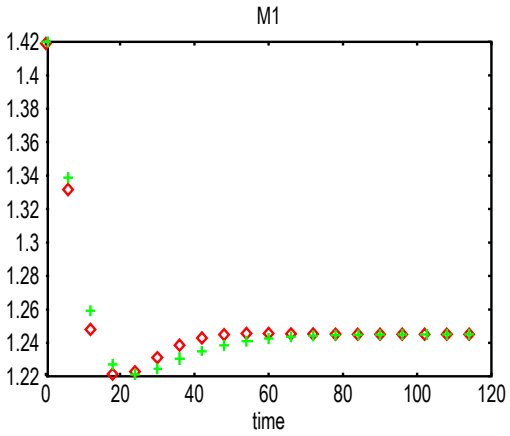
3-step model

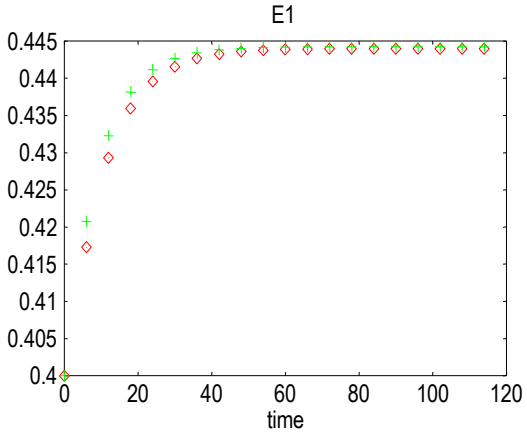
- Total of 39 parameters, 3 are equilibrium constants (independent of the biology),
- Total of 36 adjustable parameters to fit the model to the time course data
- Two classes of parameters:
  - Hill coefficients: allowed to vary [0.1:10]
  - All others: allowed to vary [10^{-12}:10^{6}]

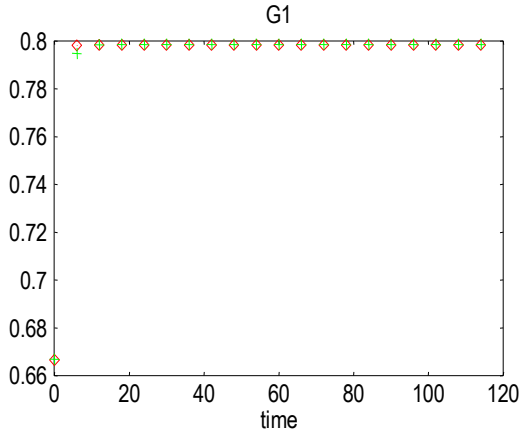
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

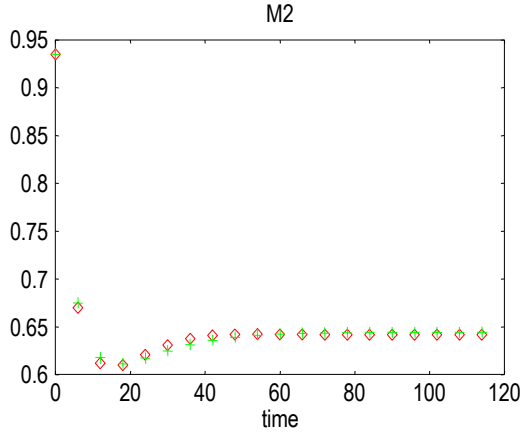
3-step model – results

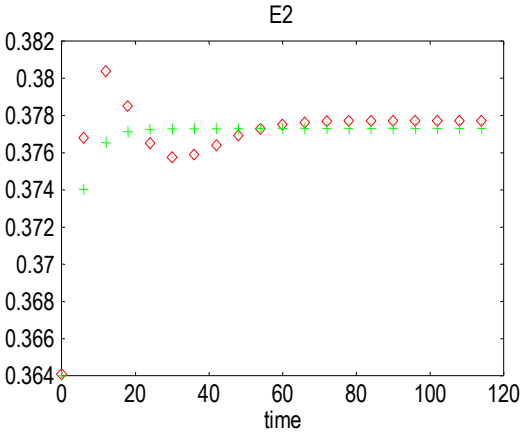
- Gradient methods could not converge to the solution from an arbitrary starting point
- Evolutionary programming converged to a small sum of squares…
- … which could be further improved with the Hooke and Jeeves method…
- …and finally with Levenberg-Marquardt
- The final solution replicates the original dynamics rather well
- Estimated parameter values sometimes deviate by 10x

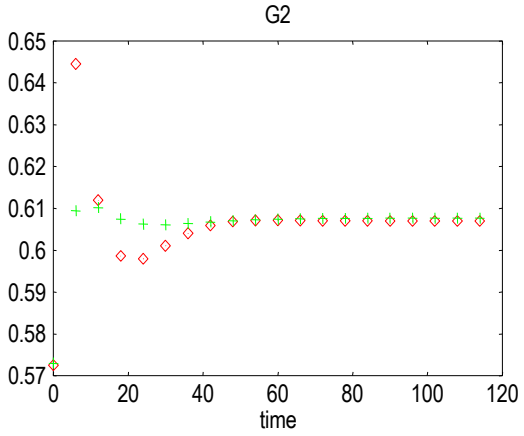


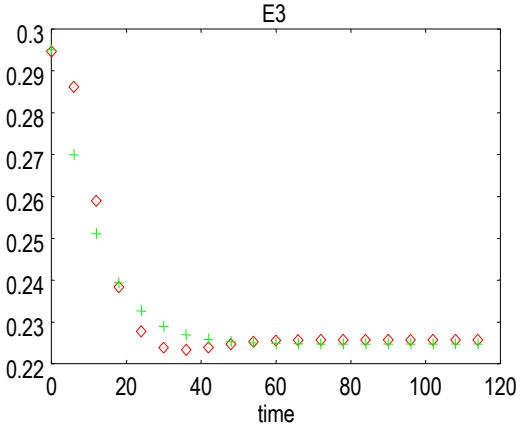


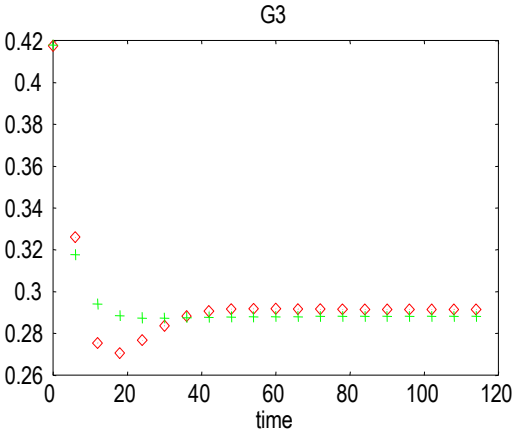






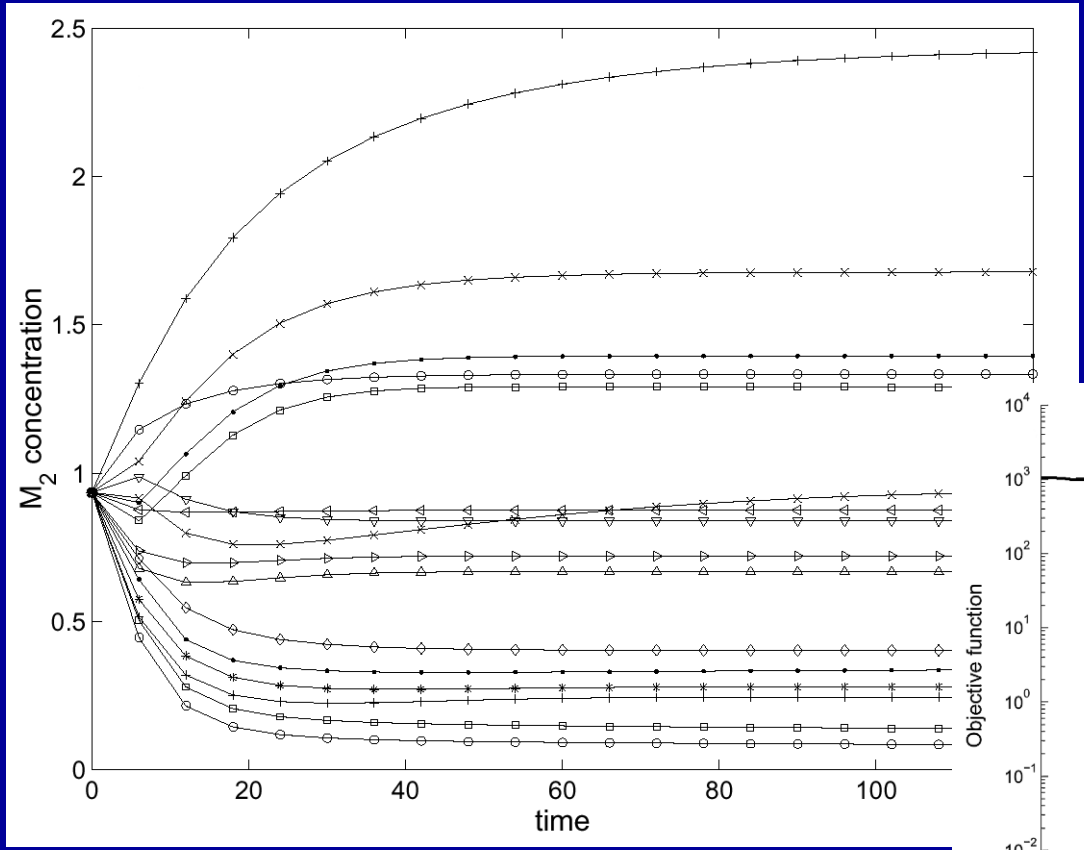




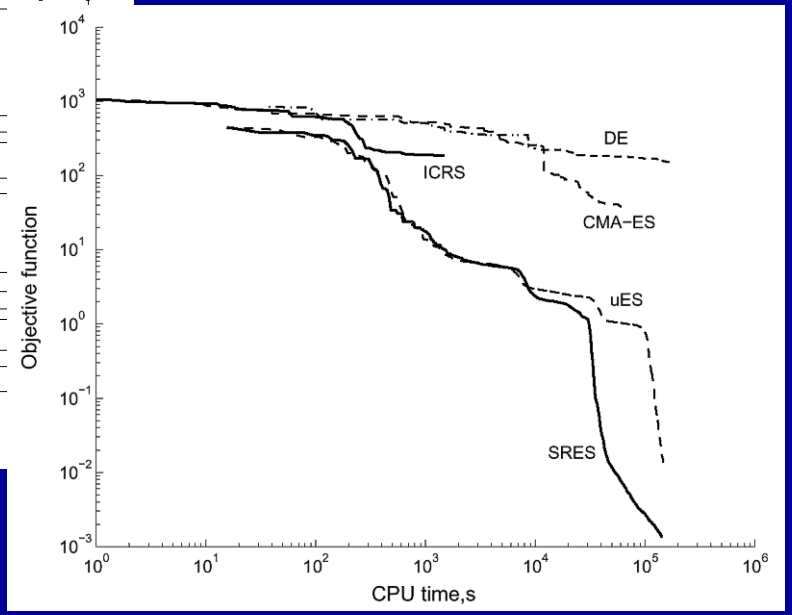


Mendes, P. (2001) Modeling large scale biological systems from functional genomic data: parameter estimation. In Kitano, H. (ed.), *Foundations of Systems Biology*. MIT Press, Cambridge, MA, pp. 163-186.

Results, 2nd attempt…



- 16 trajectories
- New minimization algorithms
- Worst parameter estimate off only by 16%



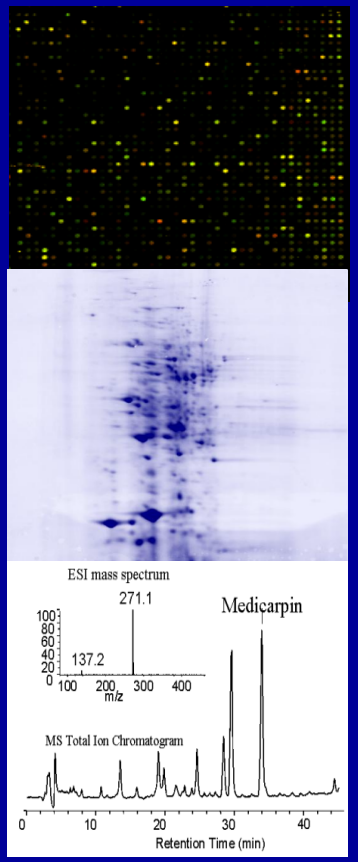
Moles, C.G., Mendes, P. and Banga, J.R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research* **13**, 2467-74.

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Issues with parameter estimation

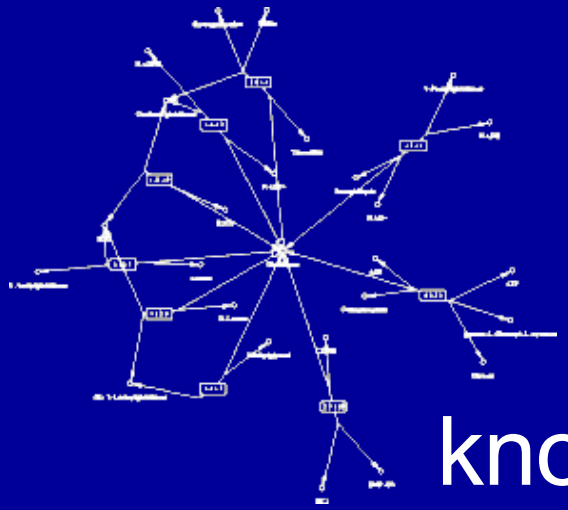
- Thousands of variables, will require extensive parallelization / distributed computing
- Functions (rhs) must be known *a priori*
- The objective function has several minima
- Data sets will often be under-determined

Top-down modeling
Reverse engineering



data

ATP	5.01
NADH	1.73
NADPH	6.21
ADP	9.91
Orthophosphate	6.7
CoA	4.15
Pyrophosphate	8.57
NH3	9.46
S-Adenosyl-L-methionine	6.24
AMP	8.49
S-Adenosyl-L-homocysteine	7.54
Pyruvate	1.22
Acetyl-CoA	4.53
L-Glutamate	1.17
2-Oxoglutarate	0.21
UDPglucose	4.83
D-Glucose	3.77
Acetate	3.4
GDP	9.55
Oxaloacetate	9.61
Glycine	4.53
L-Alanine	0.81
Succinate	0.62
UDP-N-acetyl-D-glucosamine	3.65
GTP	9
L-Lysine	6.79
Glyoxylate	0.04
L-Aspartate	6.43
Glutathione	5.52
UDP-D-galactose	0.92
Formate	7.25
L-Arginine	0.52
L-Glutamine	2.65
L-Serine	3.46
Formaldehyde	0.43
Thiamin diphosphate	9.35
Alcohol	7.61
Ascorbate	9
L-Methionine	8.8
Phosphoenolpyruvate	3.85
L-Ornithine	7.49
L-Tryptophan	6.35
L-Phenylalanine	5.07
L-Tyrosine	2.01
Malonyl-CoA	7.52
Acetaldehyde	5.28
D-Fructose 6-phosphate	6.84
Sucrose	1.29



knowledge

$$\dot{A} = \frac{V_1^f \frac{S}{K_{1S}} - V_1^r \frac{A}{K_{1A}}}{1 + \frac{S}{K_{1S}} + \frac{A}{K_{1A}}} - \frac{\left(V_2^f \frac{A}{K_{2A}}\right)\left(1 - \frac{B}{S \cdot K_{2eq}}\right)\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h-1}}{\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h} + \frac{1 + \left(\frac{C}{K_{2C}}\right)^{h}}{1 + \alpha\left(\frac{C}{K_{2C}}\right)^{h}}}$$

$$\dot{B} = \frac{\left(V_2^f \frac{A}{K_{2A}}\right)\left(1 - \frac{B}{S \cdot K_{2eq}}\right)\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h-1}}{\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h} + \frac{1 + \left(\frac{C}{K_{2C}}\right)^{h}}{1 + \alpha\left(\frac{C}{K_{2C}}\right)^{h}}} - \frac{V_3^f \frac{B}{K_{3B}} - V_3^r \frac{C}{K_{3C}}}{1 + \frac{B}{K_{3B}} + \frac{C}{K_{3C}}}$$

$$\dot{C} = \frac{V_3^f \frac{B}{K_{3B}} - V_3^r \frac{C}{K_{3C}}}{1 + \frac{B}{K_{3B}} + \frac{C}{K_{3C}}} - \frac{V_4^f \frac{C}{K_{4C}} - V_4^r \frac{P}{K_{4P}}}{1 + \frac{C}{K_{4C}} + \frac{P}{K_{4P}}}$$

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

The DREAM Project

Assessing the Accuracy of Reverse Engineering Methods

AcademyeBriefings

http://www.nyas.org/ebriefreps/main.asp?intSubsectionID=3962

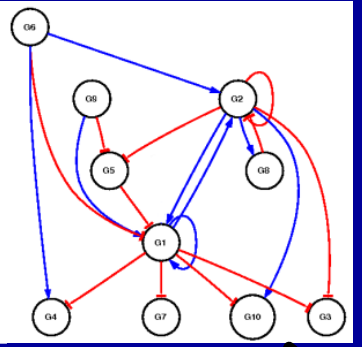
- Led by Andrea Califano (Columbia Univ.) and Gustavo Stolovitzky (IBM)
- Data and techniques to understand how well reverse engineering methods can infer the underlying biochemical networks in the cell.
- No biological gold standard available:
  - Biological data/networks (convincing but fuzzy)
  - Engineered networks (biological, yet known)
  - In silico models (full control, skepticism)
- Inspired by CASP, is expected to have a similar effect on systems biology

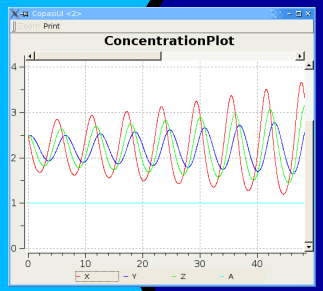
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

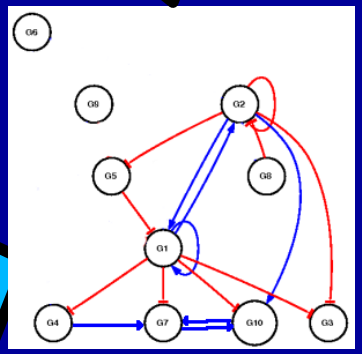
In silico biochemical networks for algorithm comparison

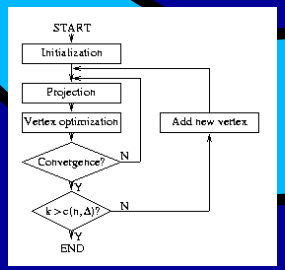
$$v_s = V \cdot \prod_j \left(\frac{Ki_j^{n_j}}{(I_j^{n_j} + Ki_j^{n_j})} \right) \cdot \prod_k \left(1 + \frac{A_k^{n_k}}{(A_k^{n_k} + Ka_k^{n_k})} \right)$$

$$v_s = V \cdot \prod_j \left(\frac{Ki_j^{n_j}}{(I_j^{n_j} + Ki_j^{n_j})} \right) \cdot \prod_k \left(1 + \frac{A_k^{n_k}}{(A_k^{n_k} + Ka_k^{n_k})} \right)$$







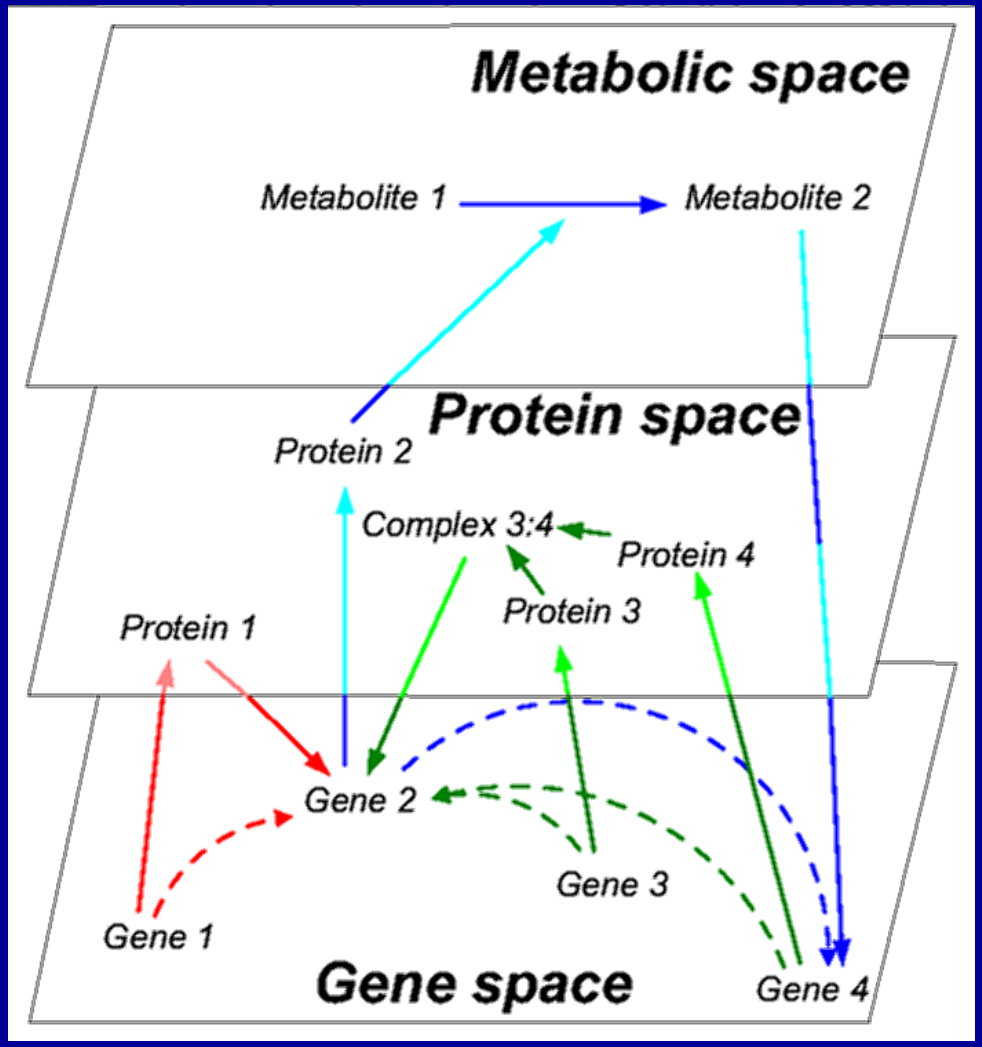


- Different experiments can be simulated
- Network is known exactly thus provides accurate metrics
- Performance of algorithms against noise can be measured
- Network details can be kept secret => blind tests

- Can these networks be made realistic enough to be relevant?

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Biochemical networks



Brazhnik, P., de la Fuente, A. & Mendes, P. (2002) *Trends in Biotechnology* **20**, 467-472.

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

BIOINFORMATICS

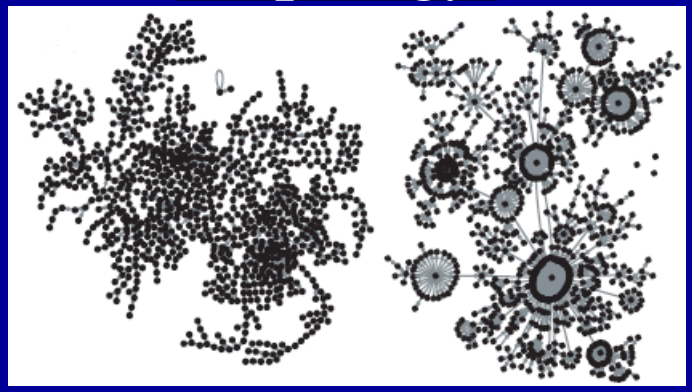
Vol. 19 Suppl. 2 2003, pages ii122–ii129
DOI: 10.1093/bioinformatics/btg1069

Artificial gene networks for objective comparison of analysis algorithms

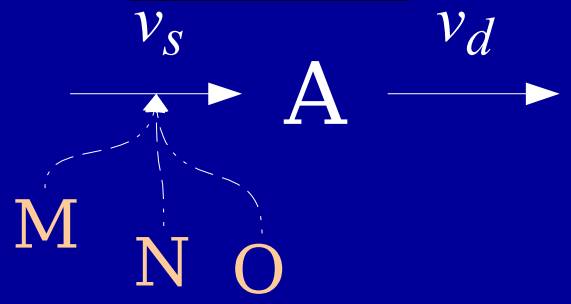
Pedro Mendes[1,*], Wei Sha[1] and Keying Ye[2]

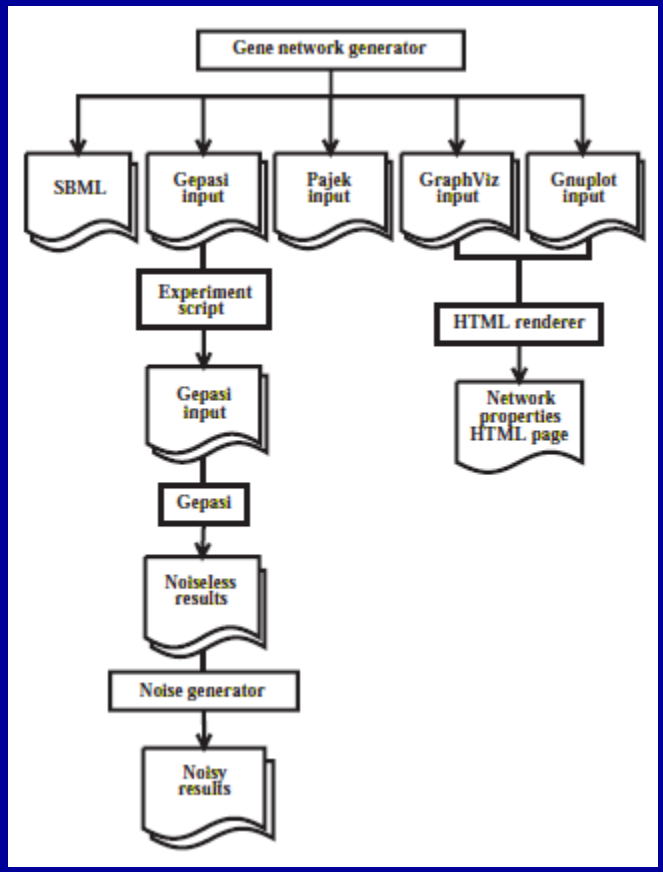
[1]Virginia Bioinformatics Institute, USA and [2]Statistics Department, Virginia Polytechnic Institute and State University, 1880 Pratt Drive, Blacksburg, VA 24061, USA

Topology



Kinetics



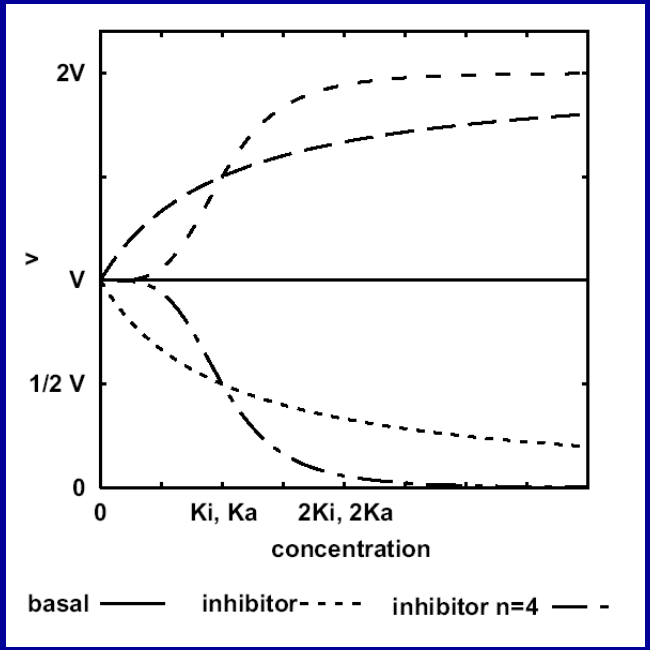


Kinetics of transcription

- Rate is saturable due to finite number of translation machinery complexes
- Rate is inhibited by a number of inhibitor genes
- Rate is activated by a number of activator genes
- Inhibition and activation can display cooperativity
- Inhibitors and activators act independently
- Rate is assumed insensitive to nucleotide concentrations

Kinetics for AGN

$$v_s = V \cdot \prod_j \left(\frac{Ki_j^{n_j}}{(I_j^{n_j} + Ki_j^{n_j})} \right) \cdot \prod_k \left(1 + \frac{A_k^{n_k}}{(A_k^{n_k} + Ka_k^{n_k})} \right)$$



- V limiting rate
- Ki inhibition constants
- ni Hill coefficient for inhibitors
- Ka activation constants
- na Hill coefficient for activators

$$v_d = k_d \cdot A$$

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Adding noise to simulated data

- It is often important to study the effect of different sources of “noise” on measurements
- Additive noise: due to measurement noise; added to simulated data at the end
- Intrinsic noise: due to time-dependent processes; injected at all points in the network continuously
- Biological variance: due to differences in cells, individuals, and cultures; added to parameters before the simulation

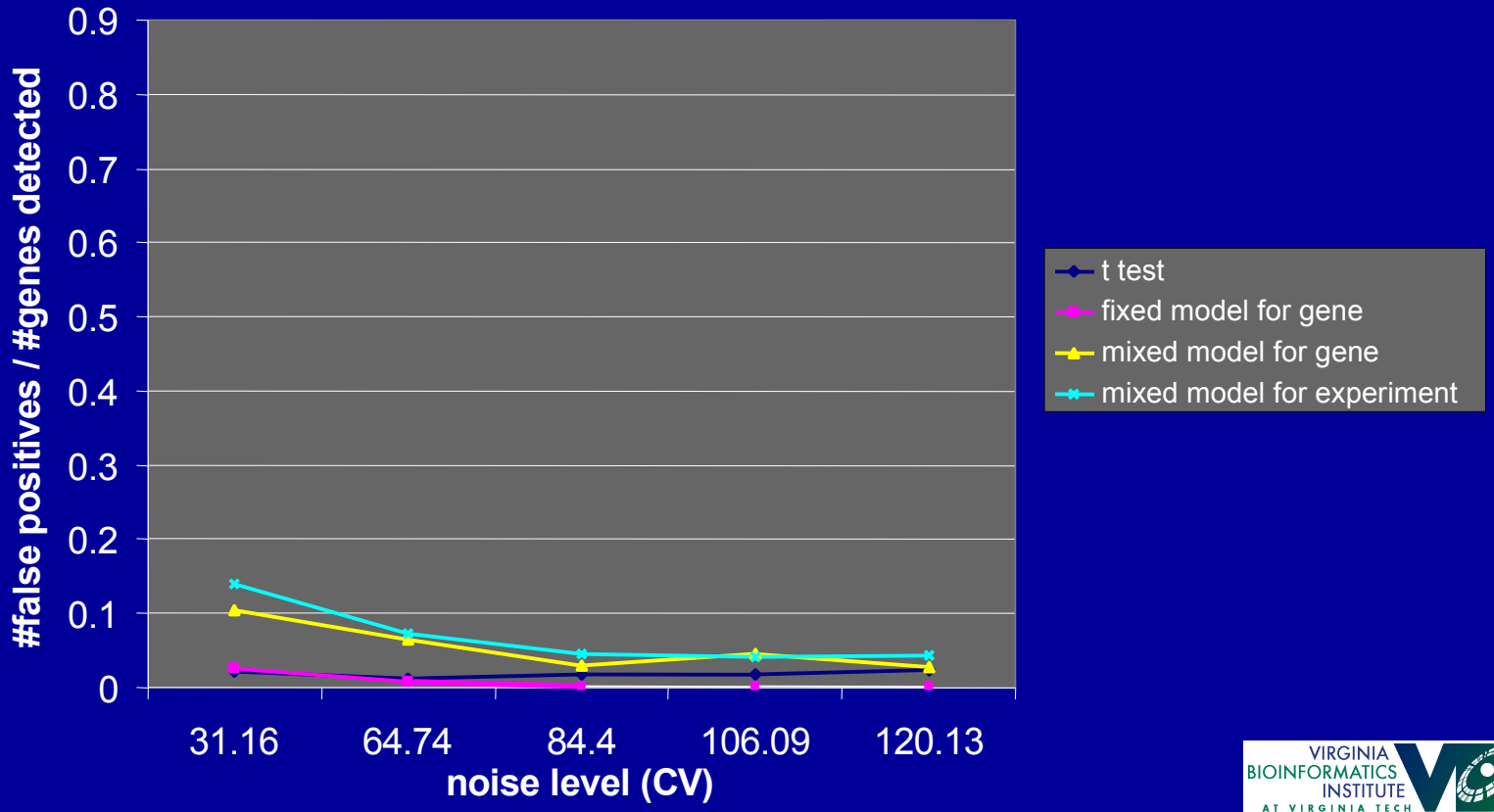
Evaluation of Statistical Methods in Microarray Differential Expression Analysis

1) Welch t test
2) ANOVA fixed model for each gene
3) ANOVA mixed model for each gene
4) ANOVA mixed model for whole experiment

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Comparison: false positives

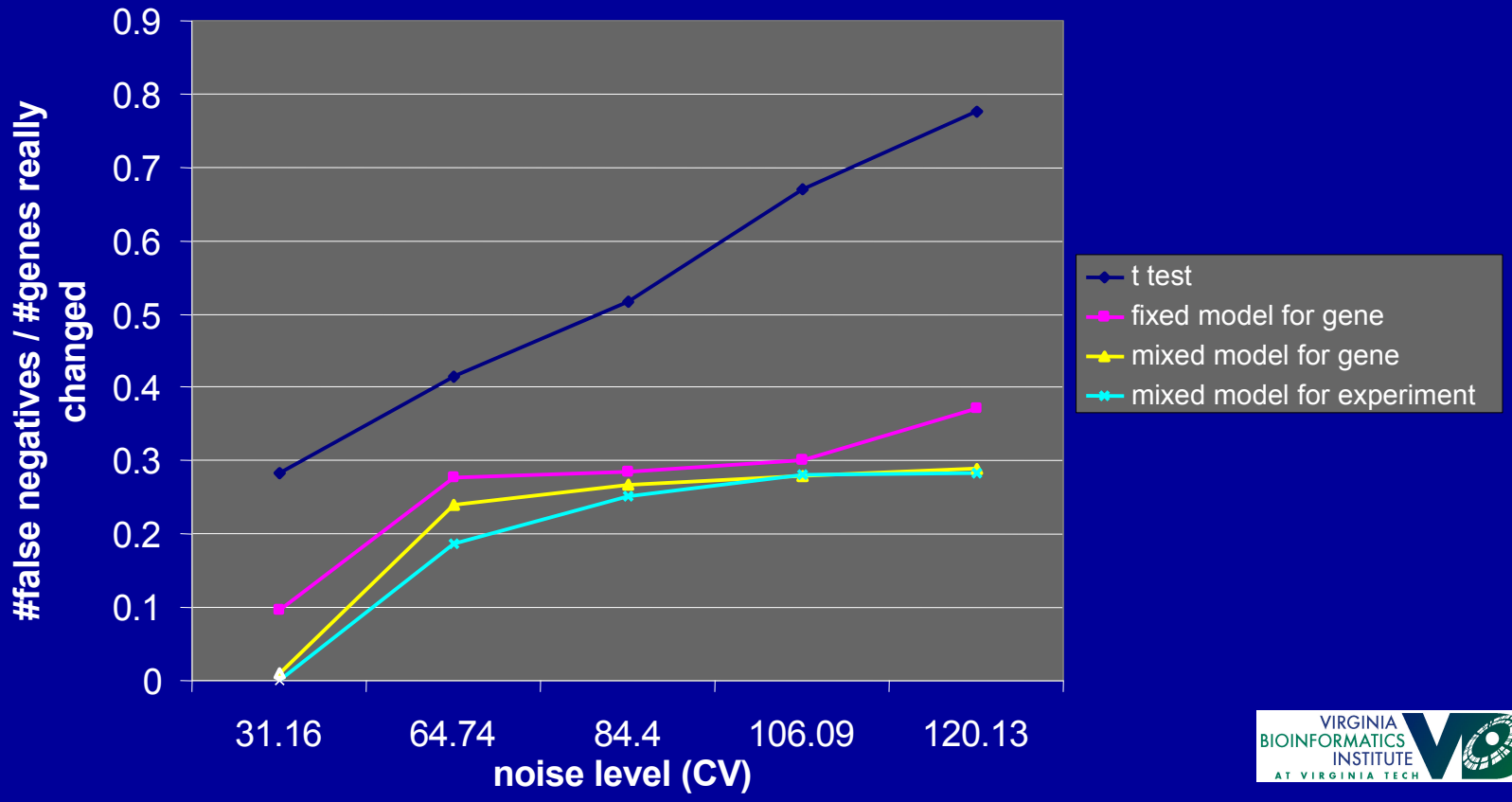
(*P* values adjusted to constant FDR of 5%)



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Comparison: false negatives

(*P* values adjusted to constant FDR of 5%)

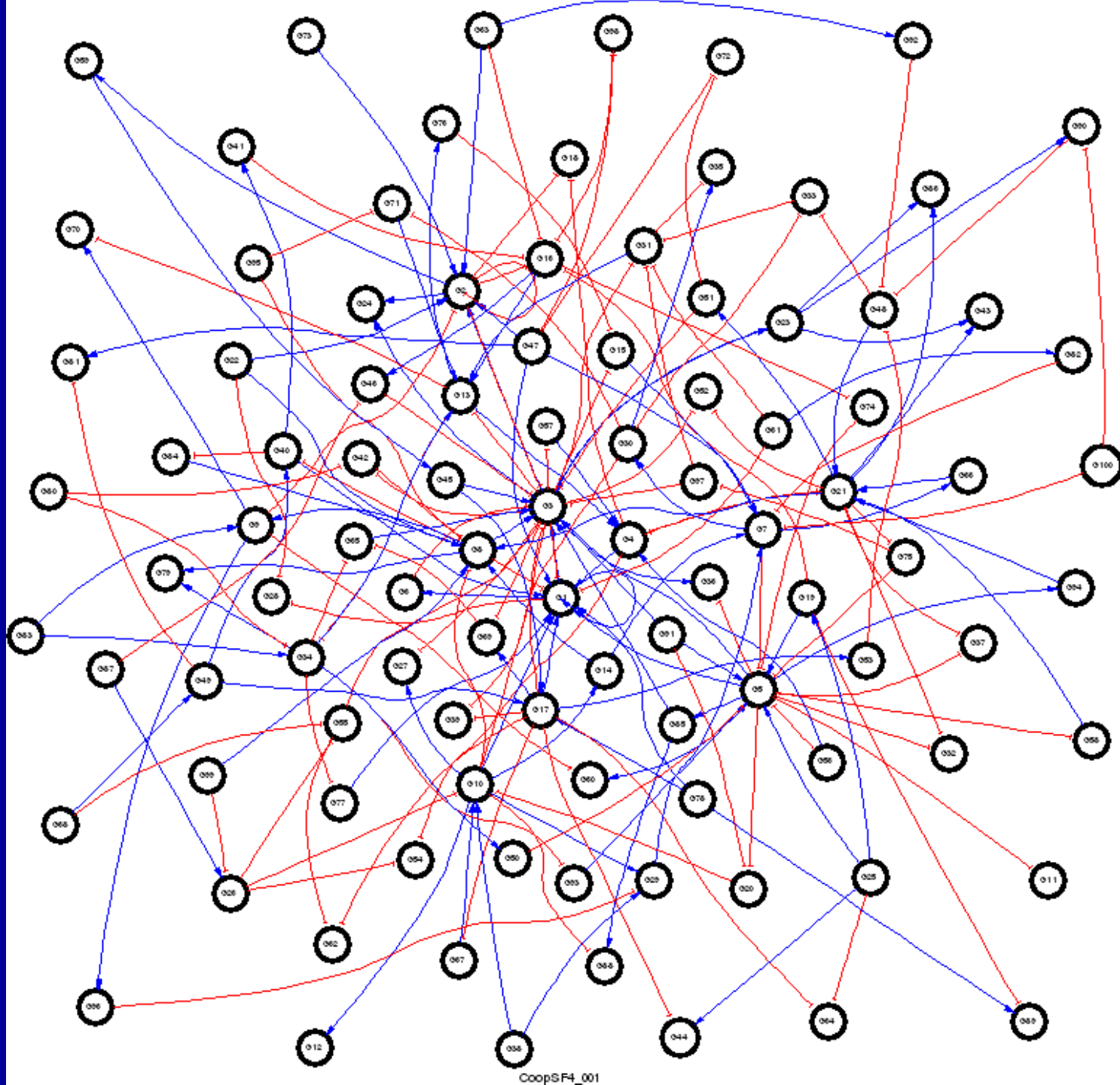


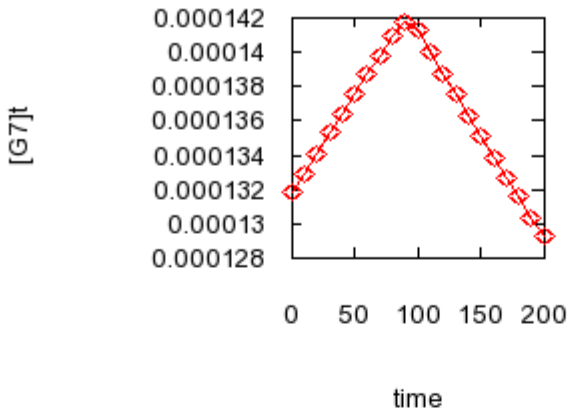
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

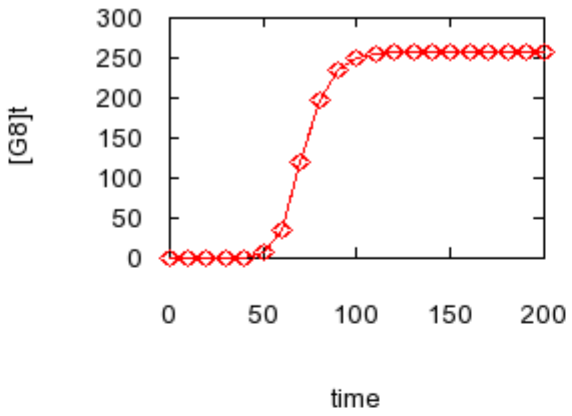
Effect of kinetics

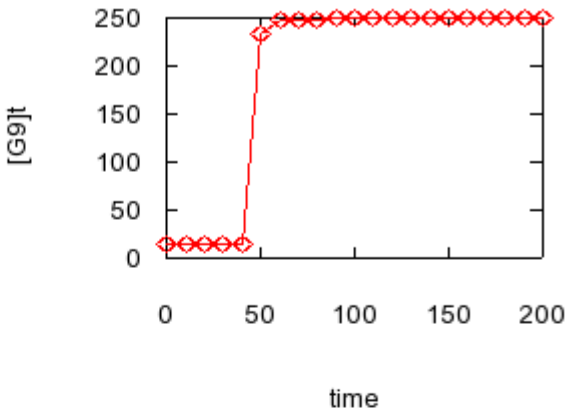
- Original networks had similar kinetics for all nodes:
  - Similar time scales
  - Similar ranges
- For a single network (CoopSF41), parameters were adjusted to obtain "interesting" behavior
  - Several time scales
  - Widely different ranges
  - Strongly nonlinear behavior

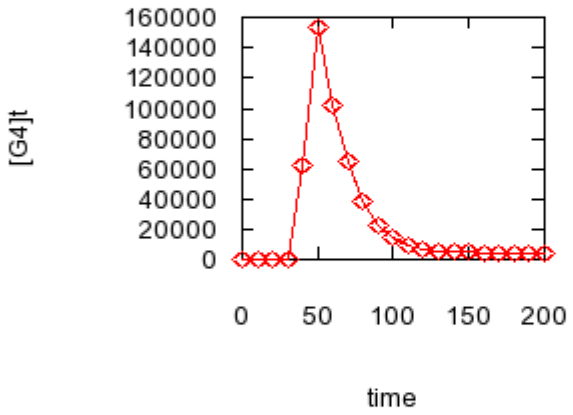
CoopSF41:

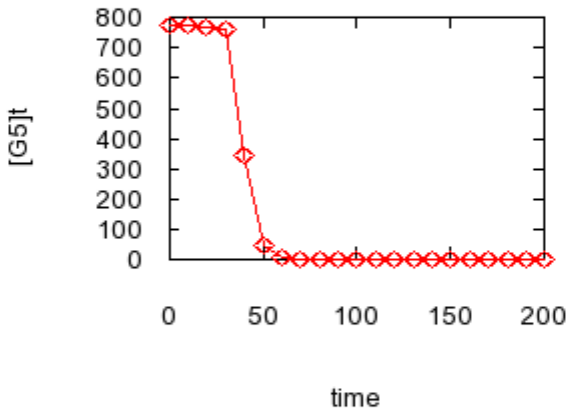


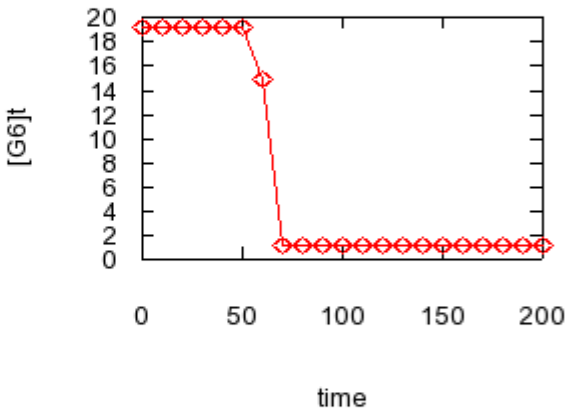

[G7]t
0.000142
0.00014
0.000138
0.000136
0.000134
0.000132
0.00013
0.000128
0
50
100
150
200
time

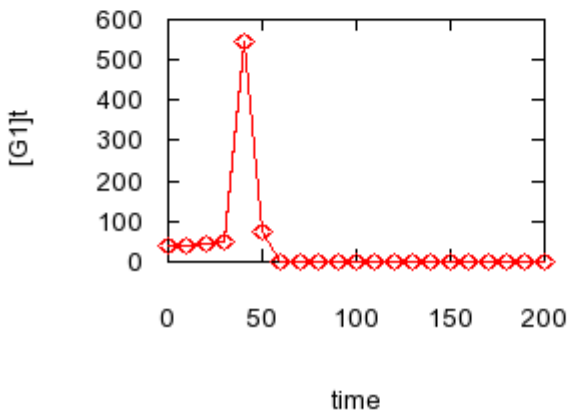

[G8]t
300
250
200
150
100
50
0
0
50
100
150
200
time

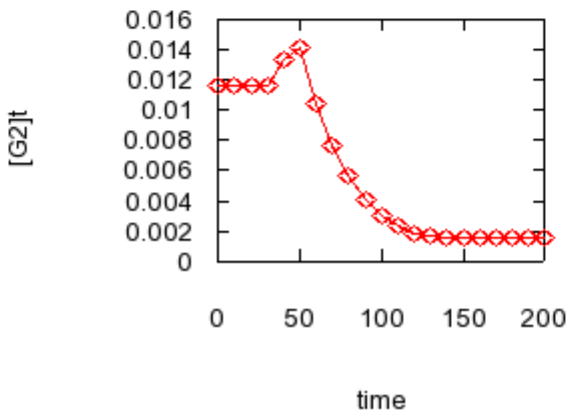

[G9]t
250
200
150
100
50
0
0
50
100
150
200
time

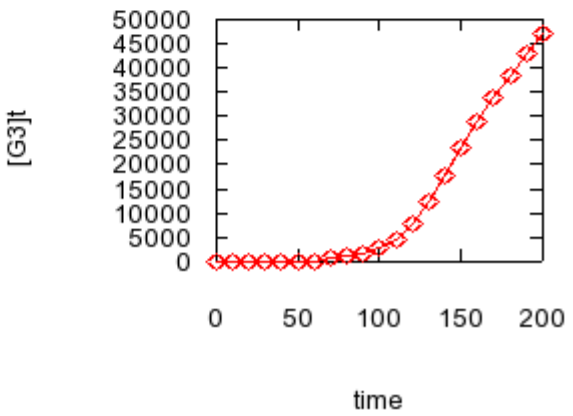

[G4]t
160000
140000
120000
100000
80000
60000
40000
20000
0
0
50
100
150
200
time


[G5]t
800
700
600
500
400
300
200
100
0
0
50
100
150
200
time


[G6]t
20
18
16
14
12
10
8
6
4
2
0
0
50
100
150
200
time


[G1]t
600
500
400
300
200
100
0
0
50
100
150
200
time


[G2]t
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
time


[G3]t
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
time

Metrics

Confusion matrix

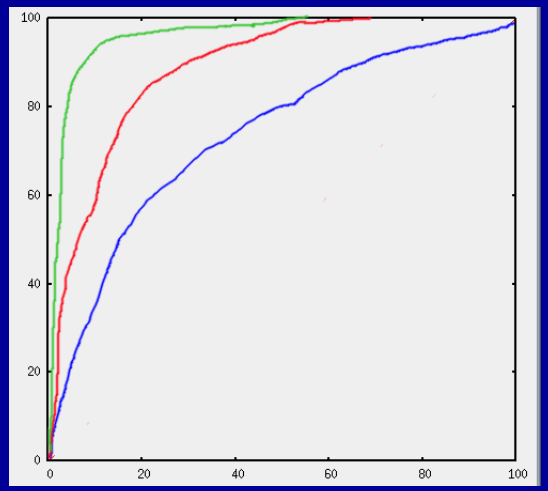
- Measures biases on inference of connections

predicted

real	+	0	-	
+	7	2	1	70.00%
0	3	10	2	66.67%
-	1	2	8	72.73%
	63.64%	71.43%	72.73%	69.44%

ROC curves

- Effect of algorithm tuning parameters on specificity and sensitivity



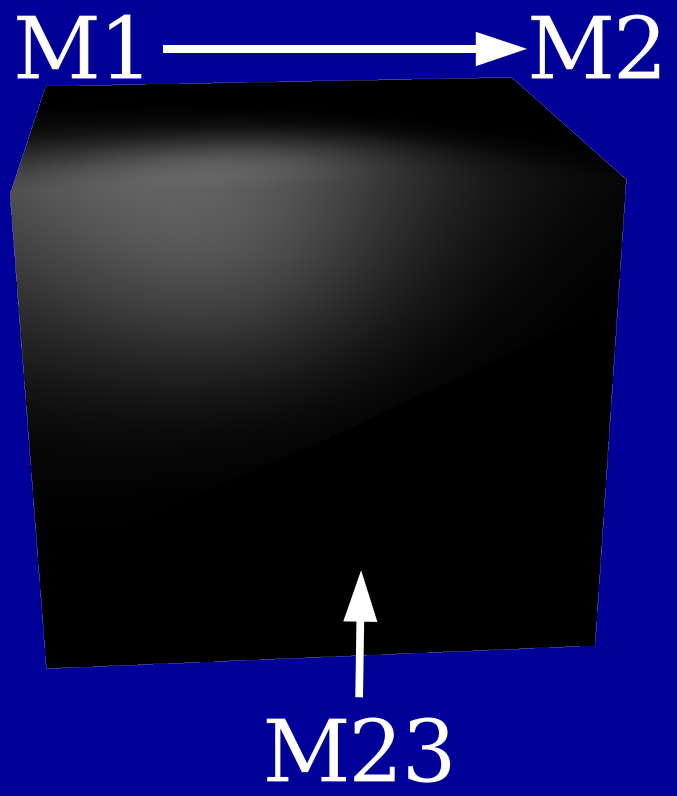
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

How about the rest of biochemistry ?

- To approach what happens in real cells, we need to include metabolism and signaling
- Metabolism is a very rigid network
  - A few cofactors link many reactions (ATP, NADP, CoA)
  - There are branches almost at every metabolite
- Signaling transduction is very flexible
  - Large crosstalk between pathways
  - Integration of signals and combinatorial responses

The Claytor Network

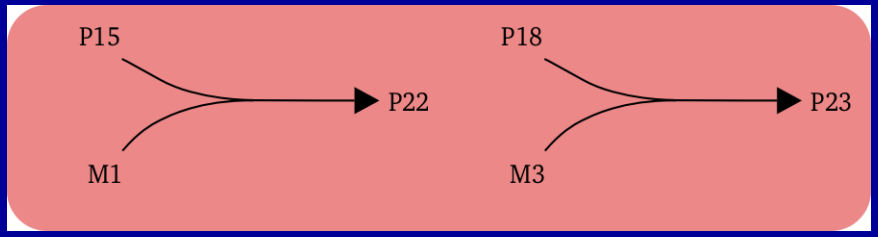
- Transcription
- Translation
- Metabolism
- Signal transduction
- 20 genes
- 23 protein forms
- 16 metabolites
- 3 transcription factors
- 2 receptors



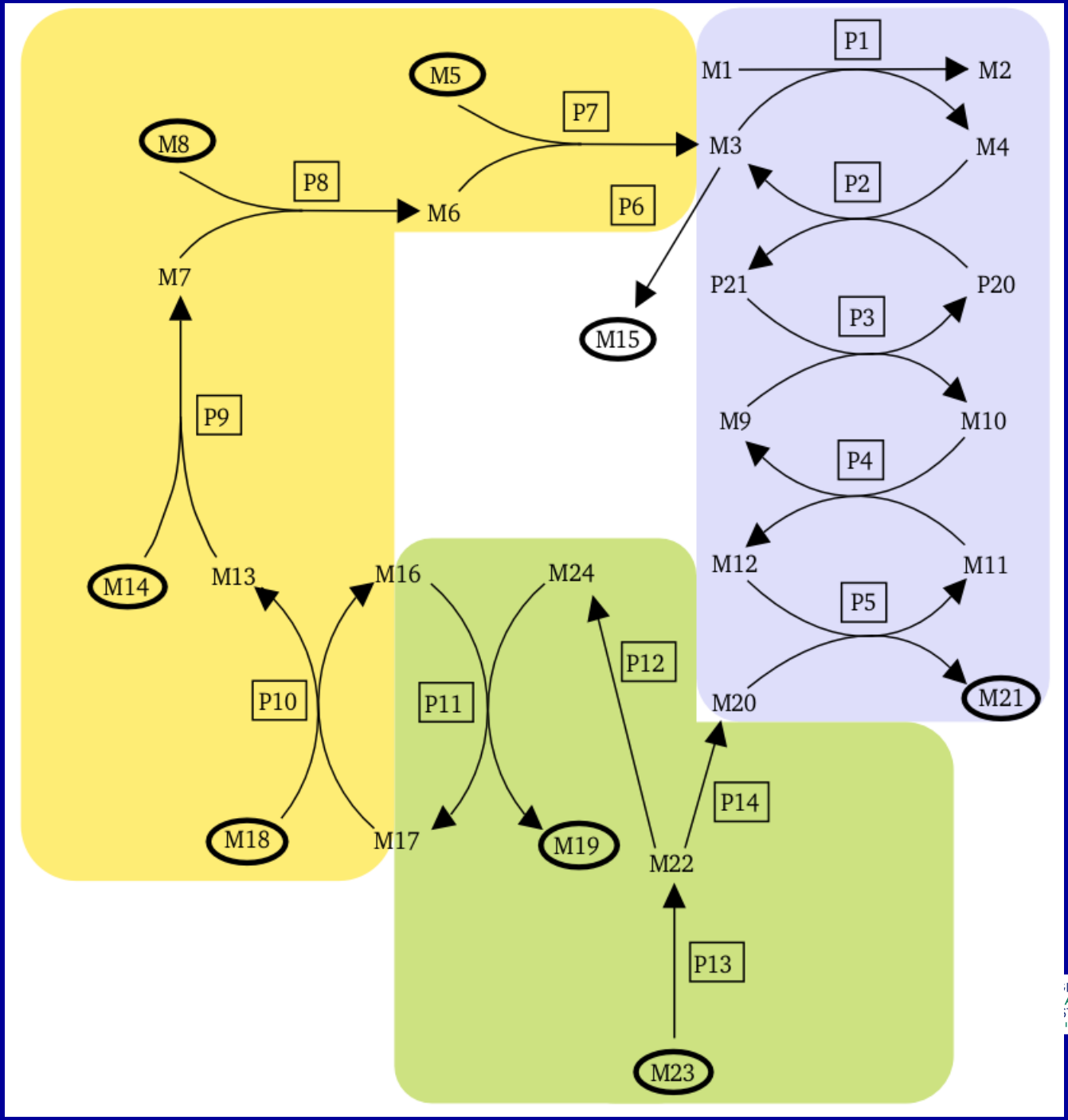
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Protein synthesis and signaling

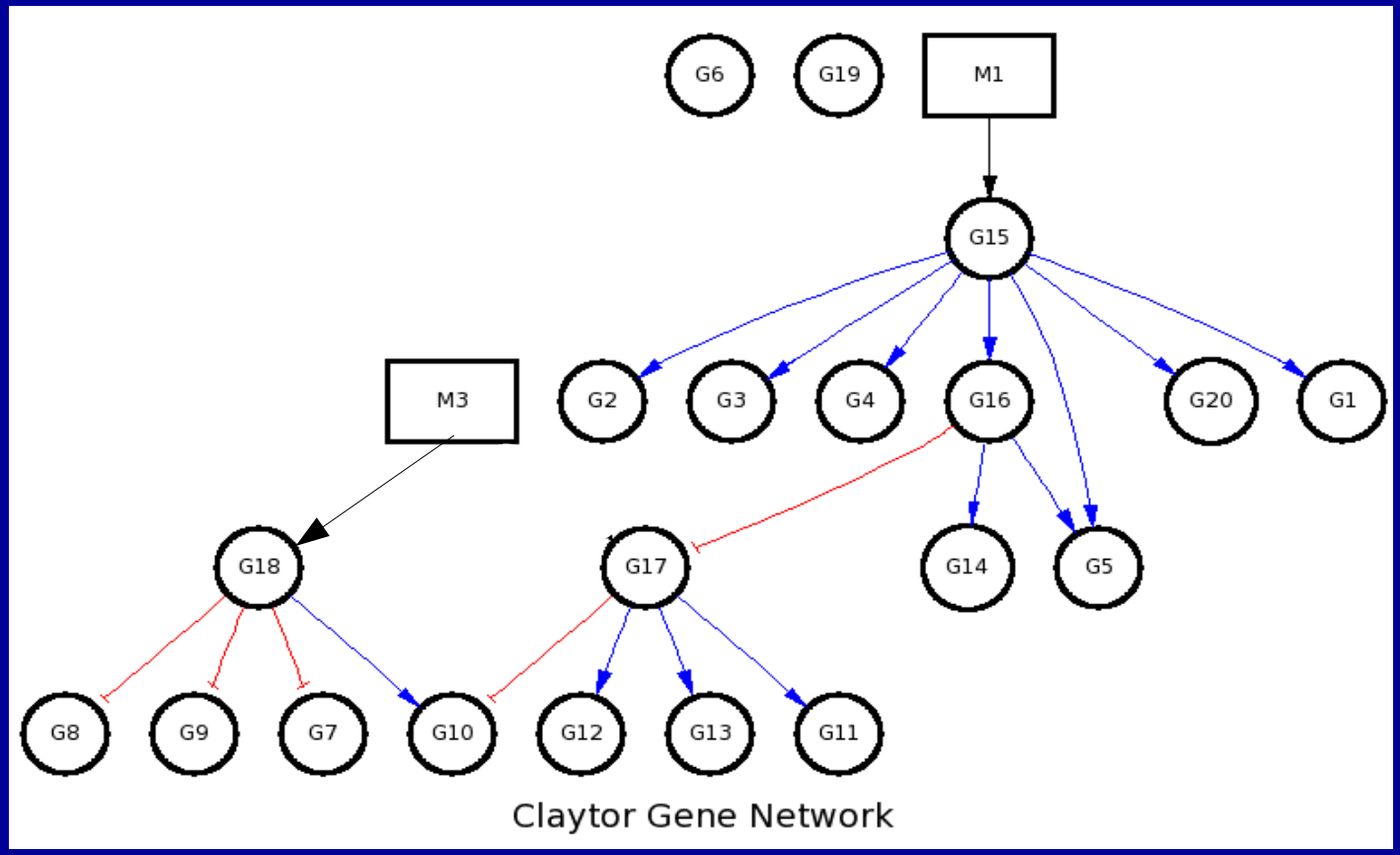
G1 → P1
G2 → P2
G3 → P3
G4 → P4
G5 → P5
G6 → P6
G7 → P7
G8 → P8
G9 → P9
G10 → P10
G11 → P11
G12 → P12
G13 → P13
G14 → P14
G15 → P15
G16 → P16
G17 → P17
G18 → P18
G19 → P19
G20 → P20



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI


M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
M21
M22
M23
M24
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P20
P21

Genetic regulation

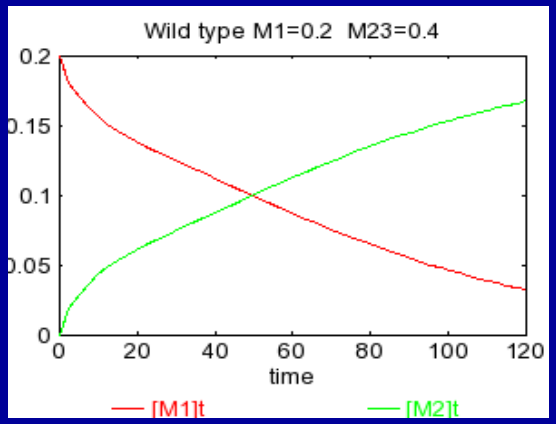


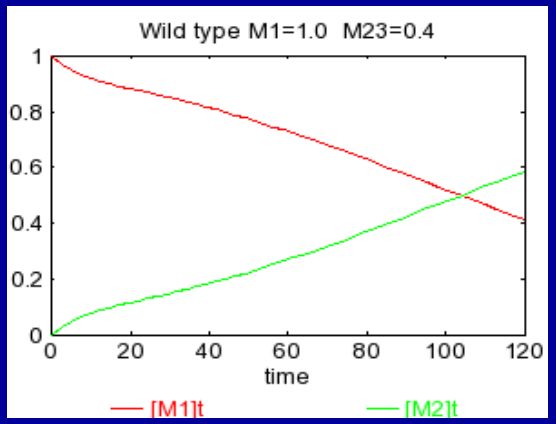
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

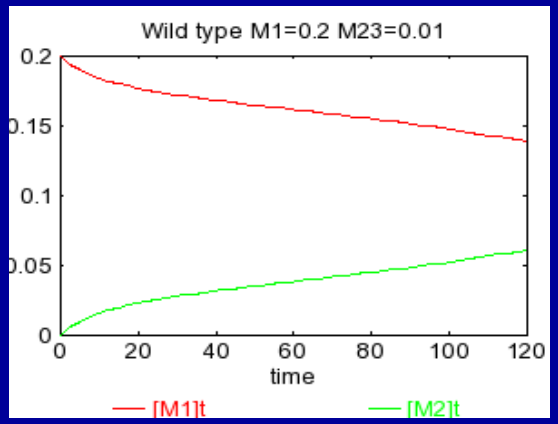
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





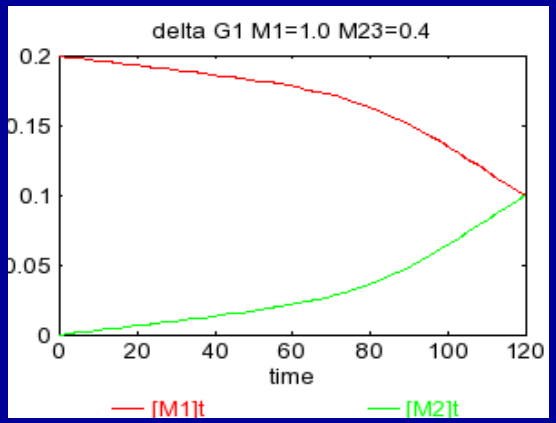


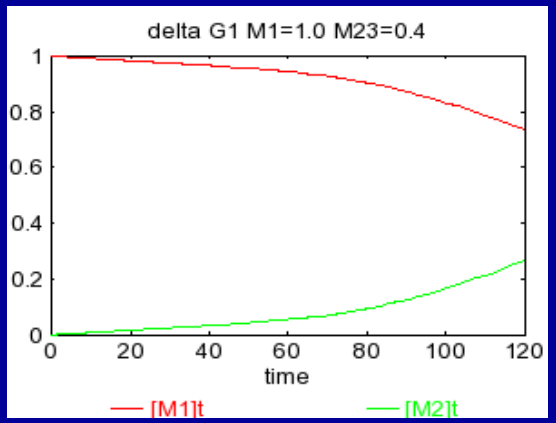
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

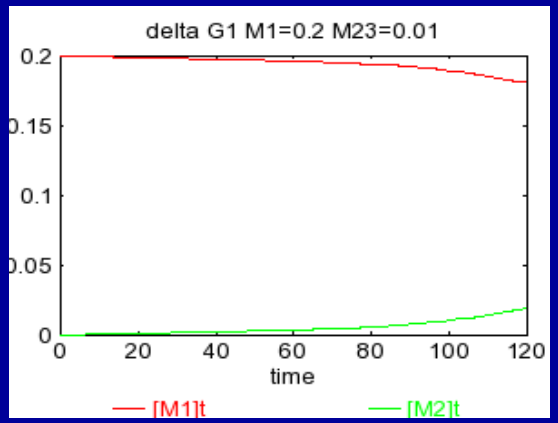
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





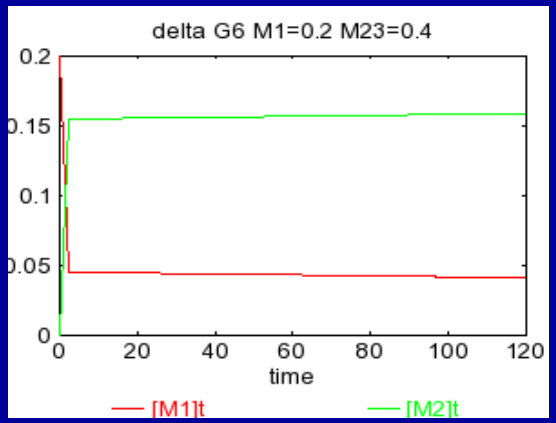


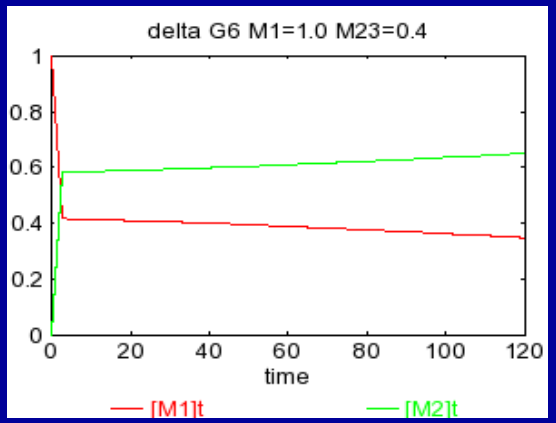
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

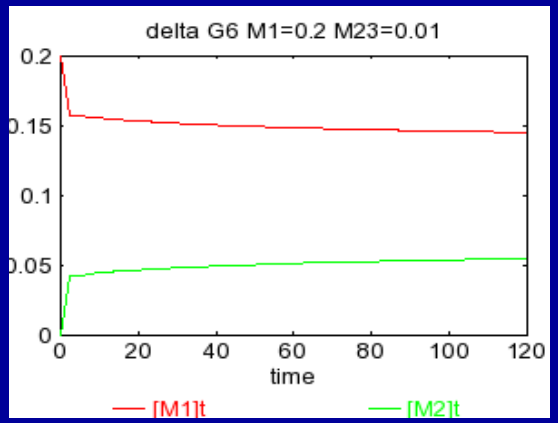
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





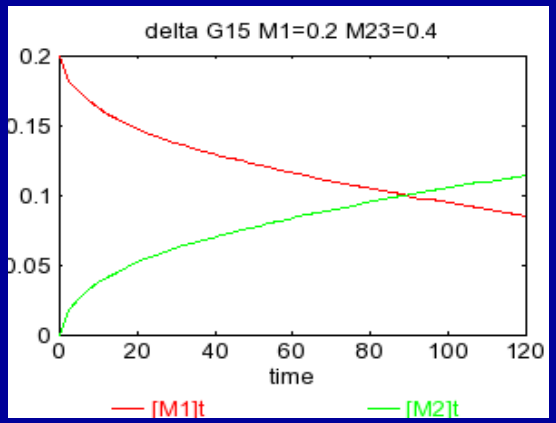


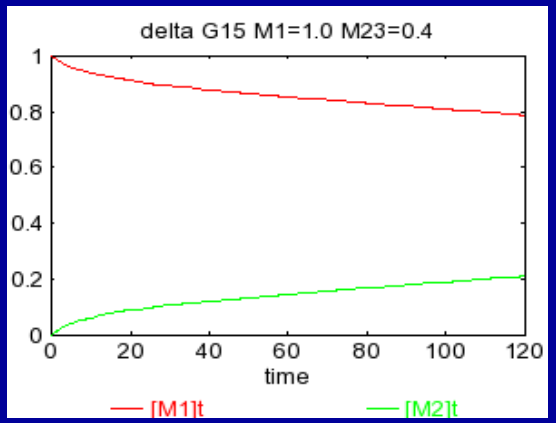
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

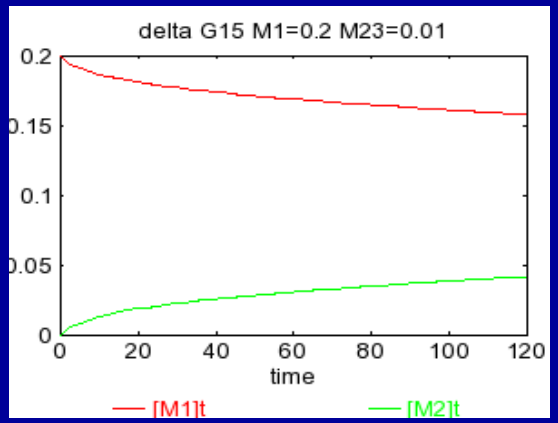
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





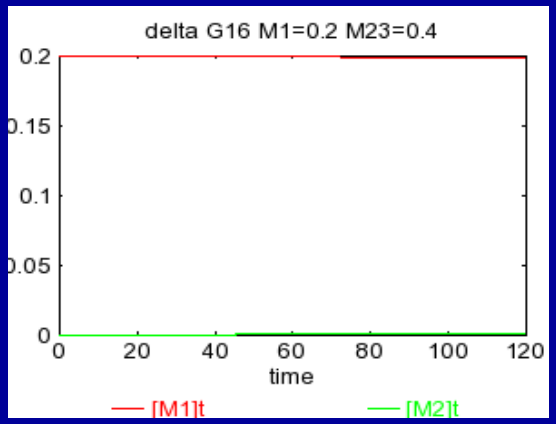


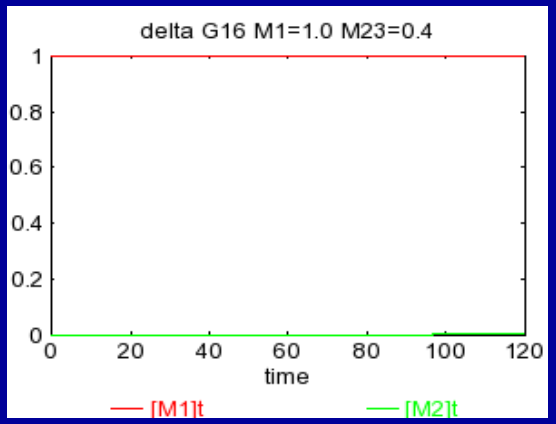
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

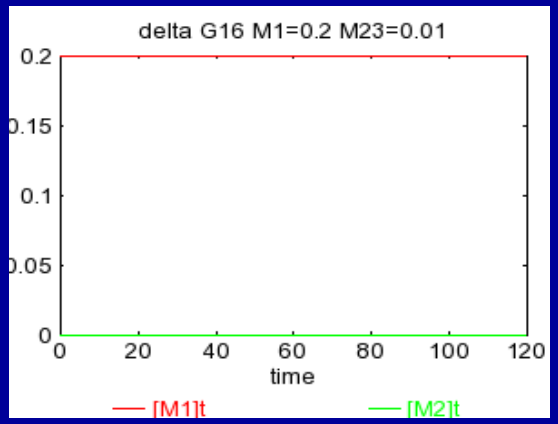
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





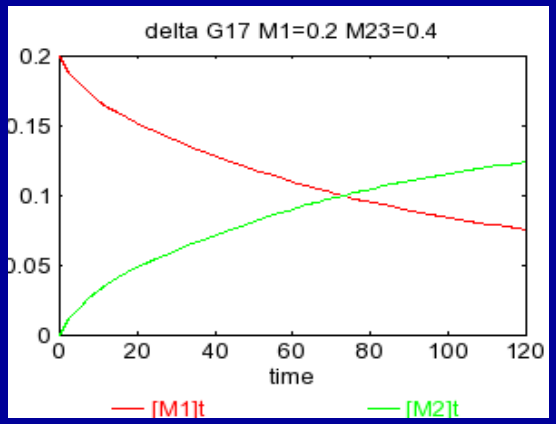


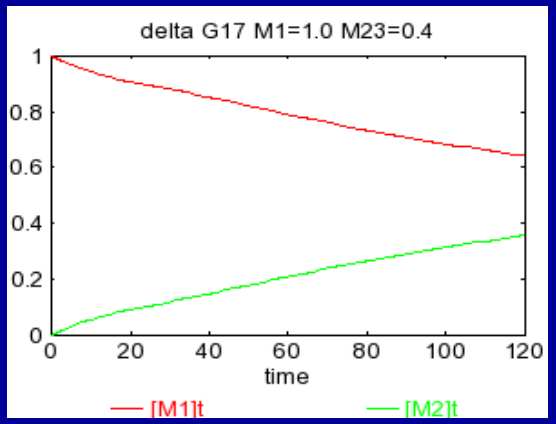
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

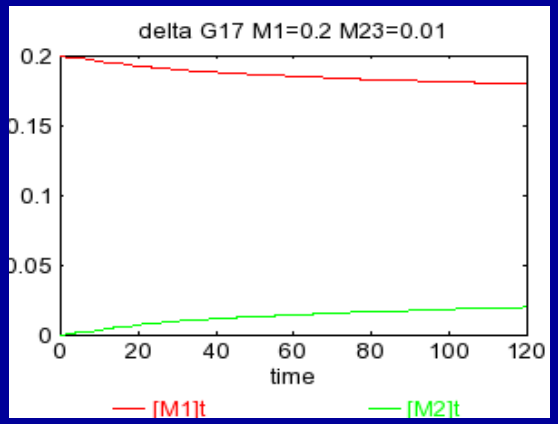
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





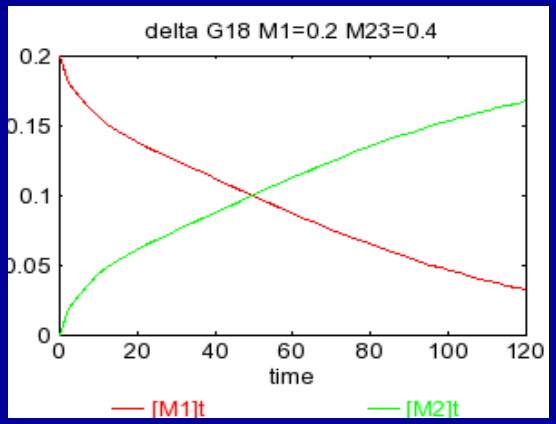


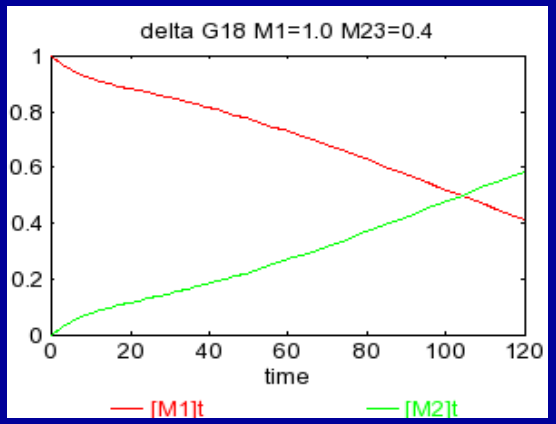
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

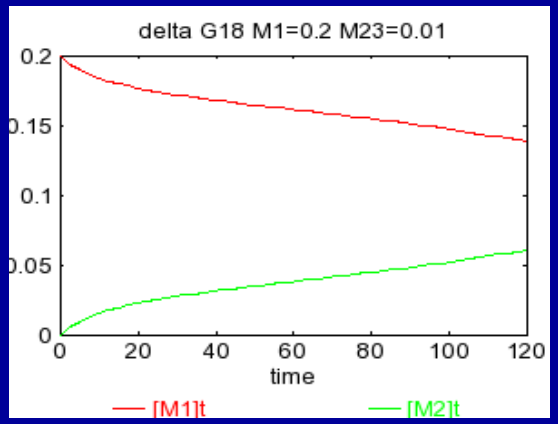
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19





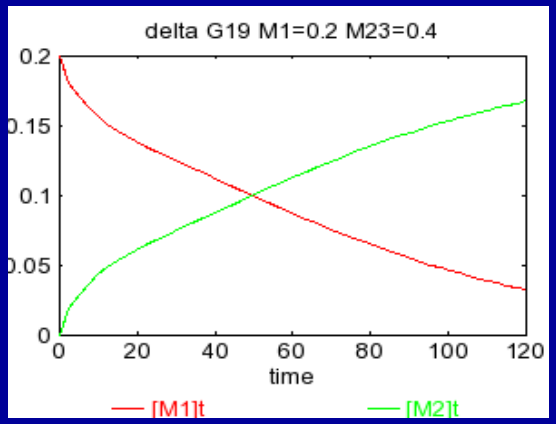


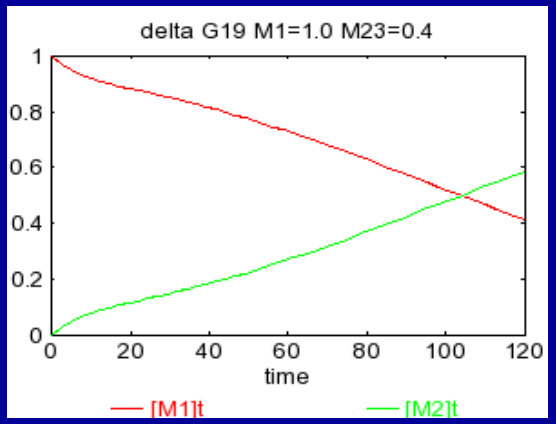
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

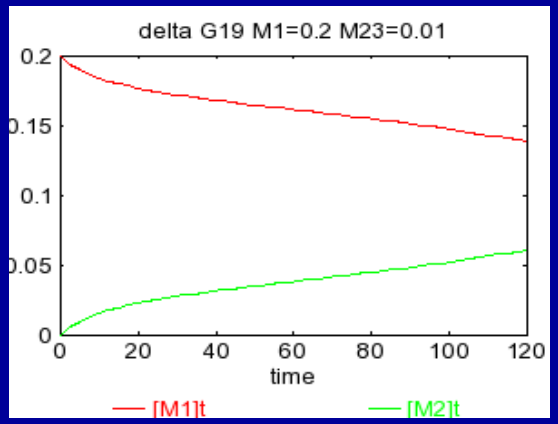
Gene knock-out experiments

STRAINS

- Wild type
- ΔG1
- ΔG6
- ΔG15
- ΔG16
- ΔG17
- ΔG18
- ΔG19







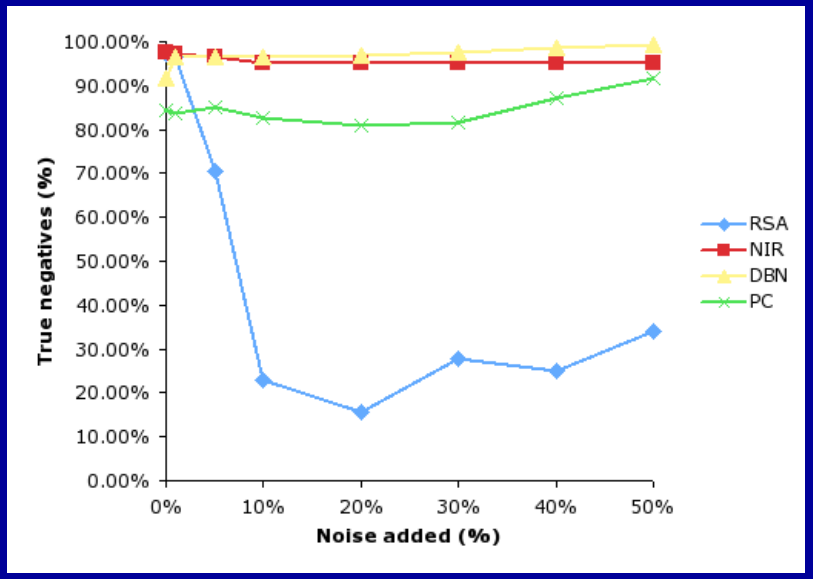
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

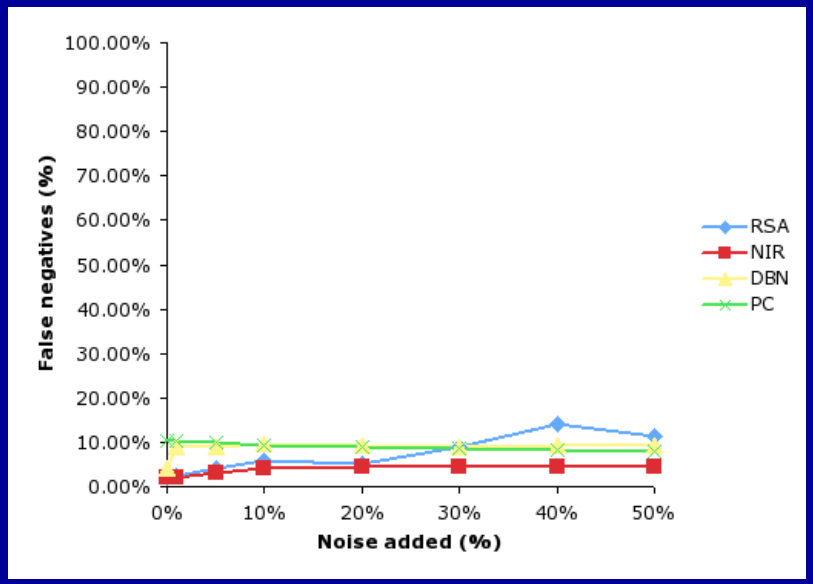
Comparison of four reverse engineering methods

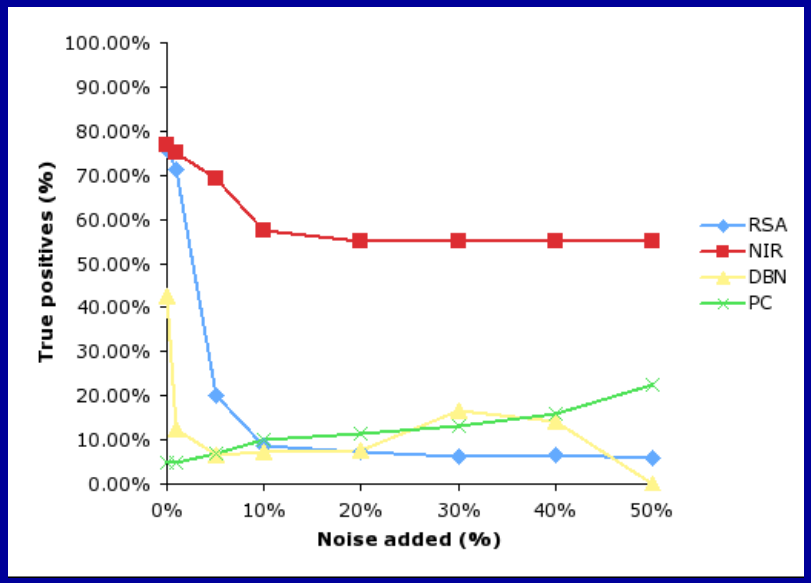
- RSA: de la Fuente, A., Brazhnik, P. & Mendes, P. (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genetics* **18**, 395-398
- NIR: Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling Science **301**, 102-105
- DBN: Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics **20**, 3594-603
- PC: de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565-3574

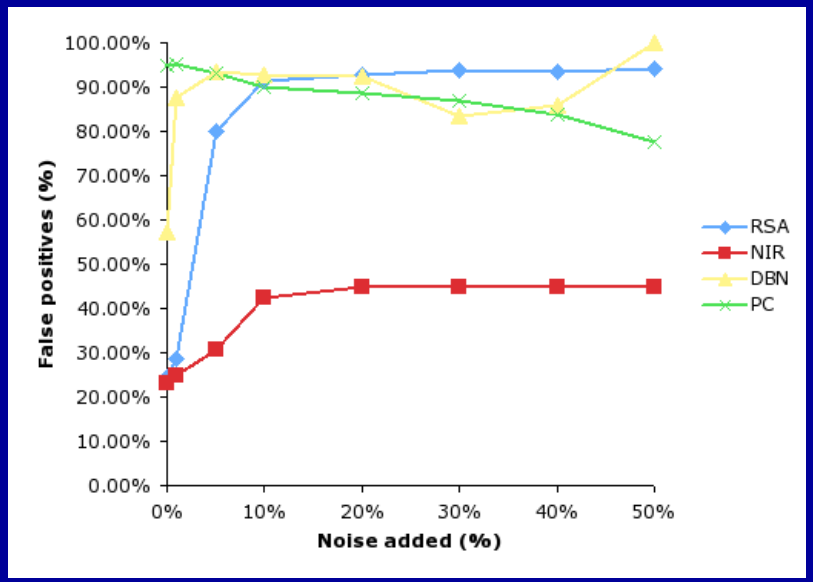
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Increasing noise levels

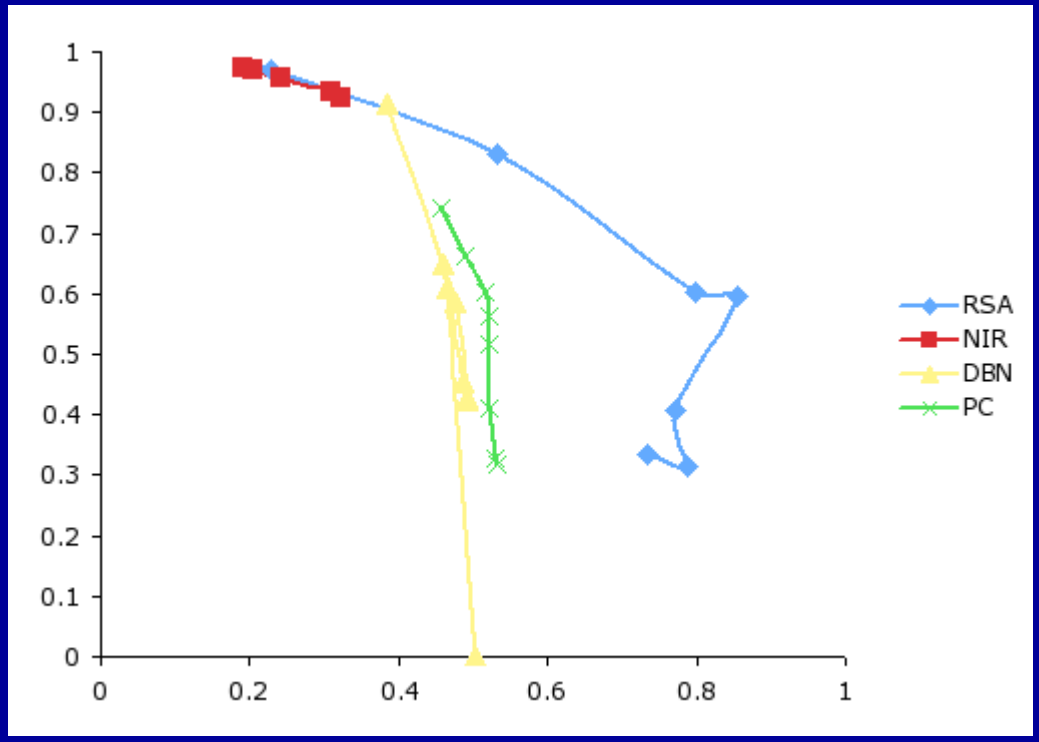








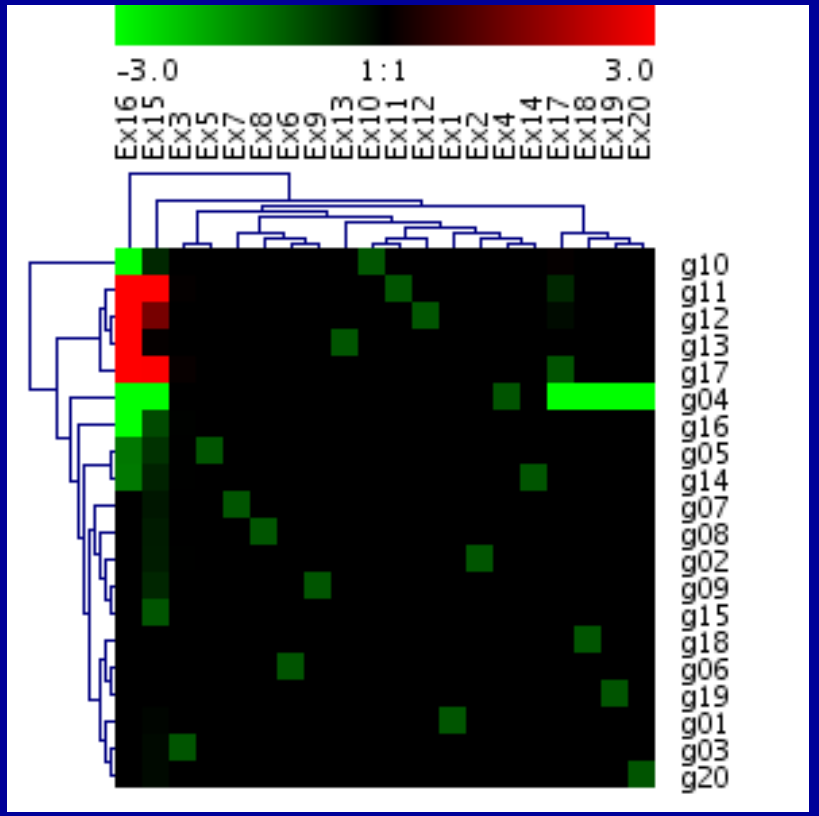
ROC curve for effect of noise



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

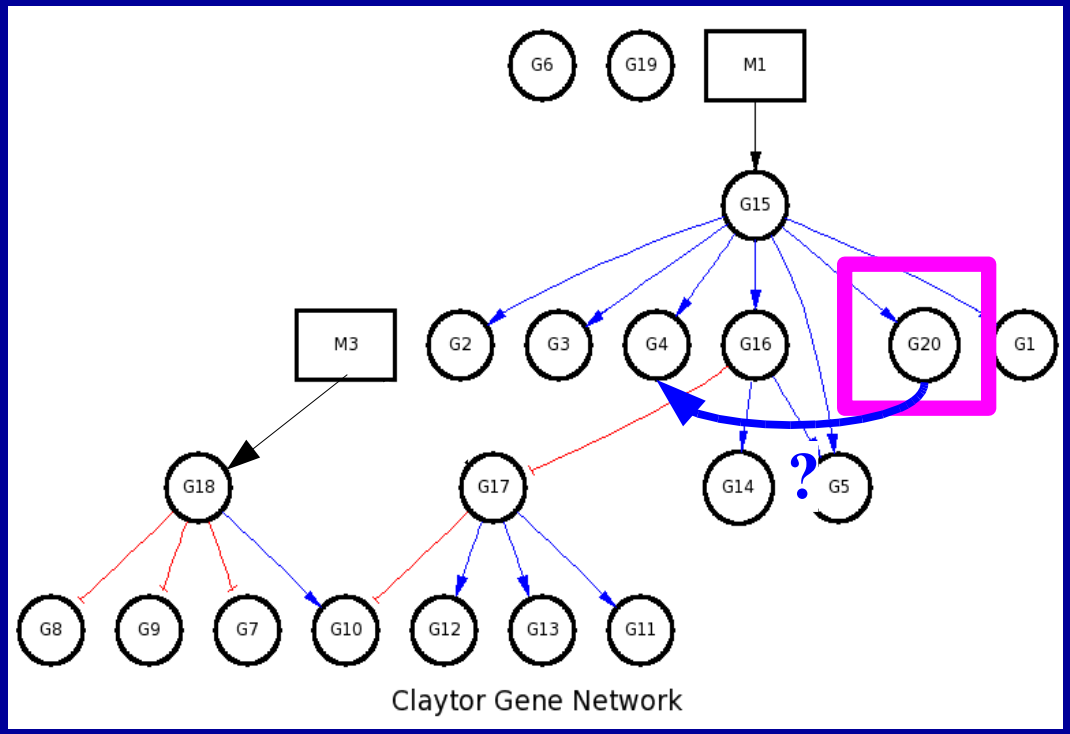
Heterozygous mutant experiments

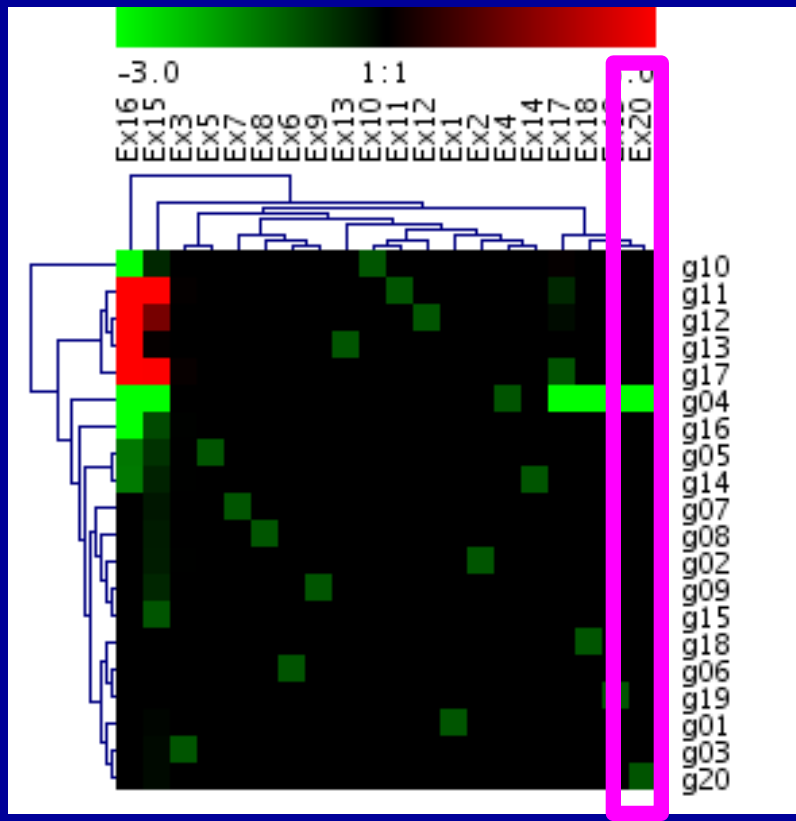
- Each gene is mutated to have 50% translation rate
- Steady-state levels of all genes are measured and represented as ratio to *wt*
- Proteins and metabolites are not measured
- REALISM:
  - Response is a result of full biochemical network
  - Most mutations have little effect



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

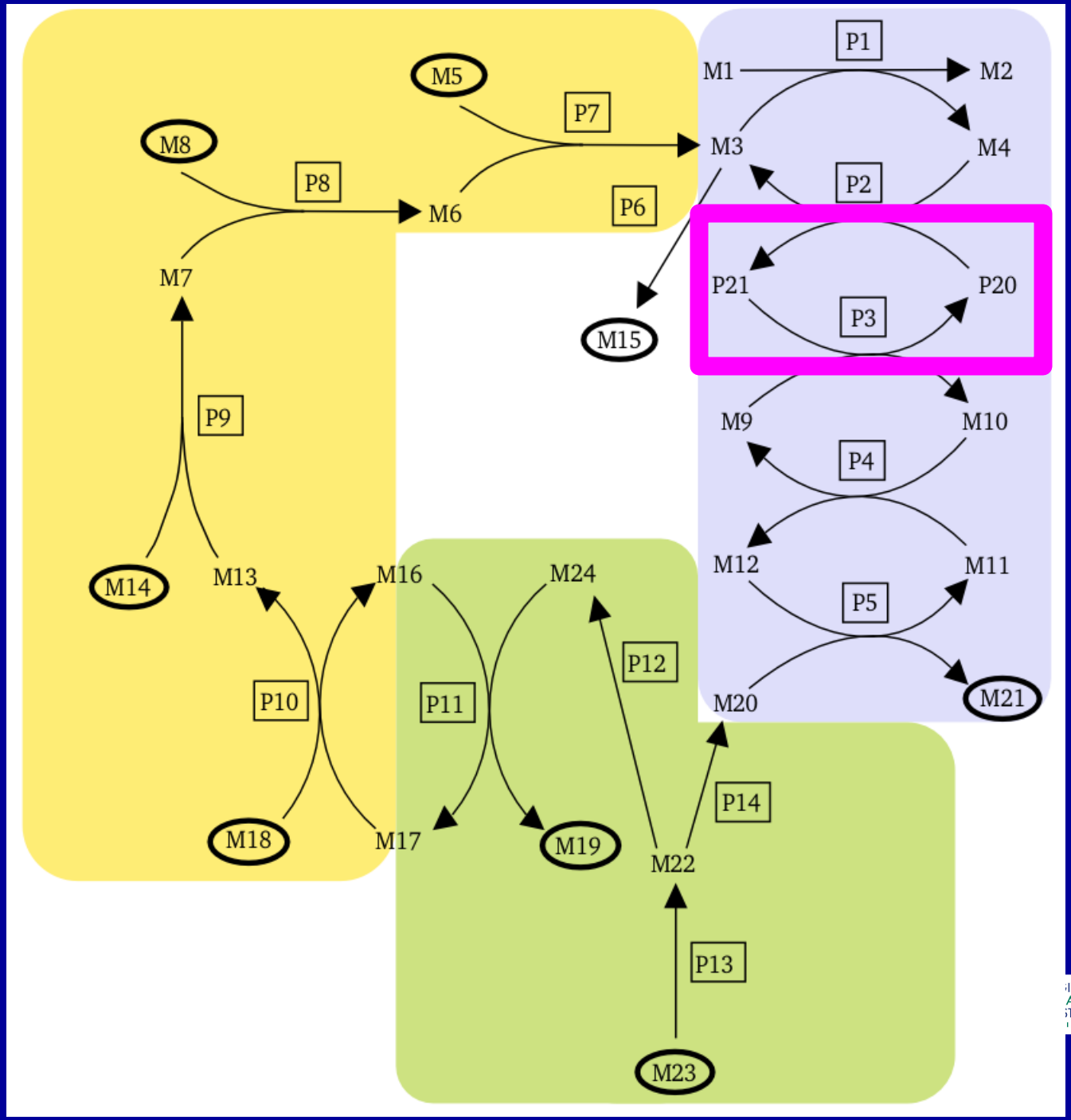
But do we really **know** the gene network accurately?





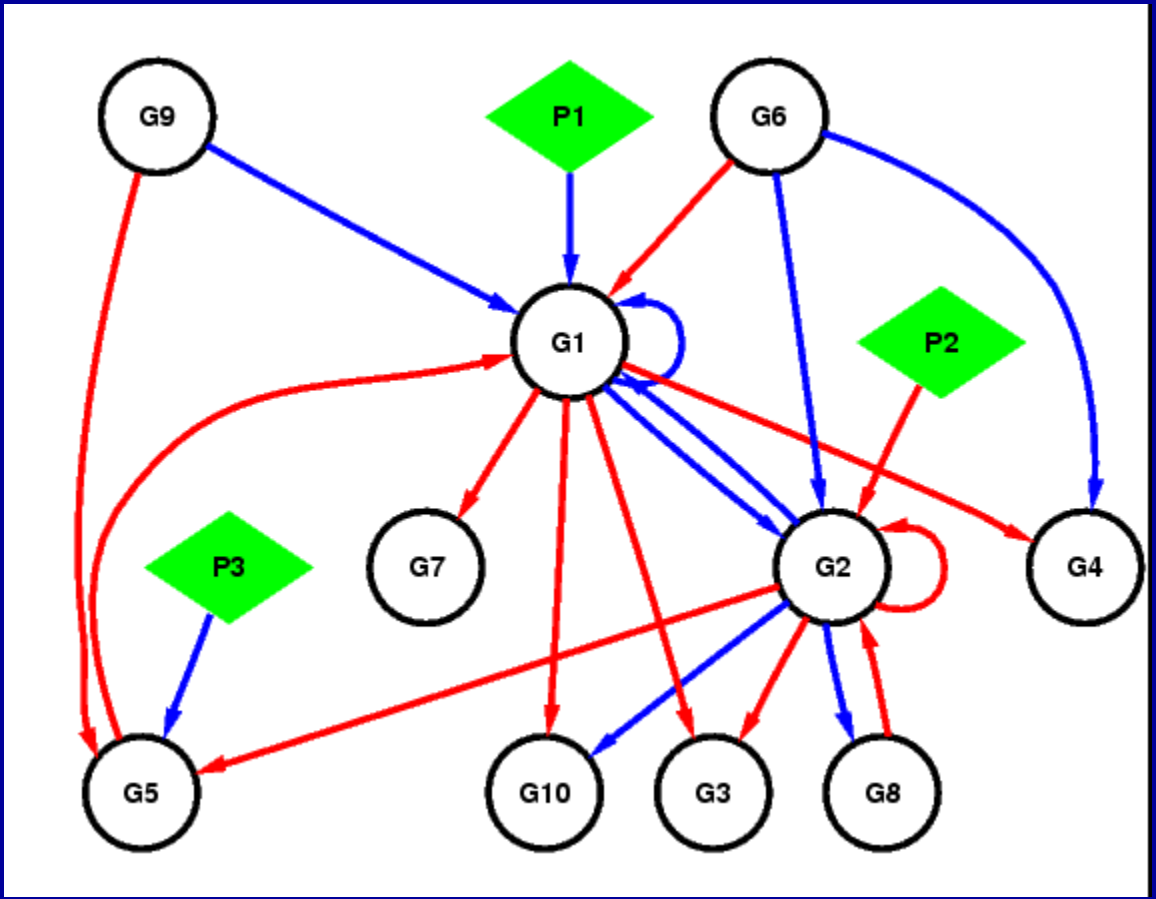
- Failure to identify all "connections" between genes
- No longer possible to calculate metrics accurately

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI


M5
P7
M8
P8
M6
P6
M7
P9
M14
M13
M16
M24
P10
P11
P12
M18
M17
M19
M22
P14
M23
P13
P1
M1
M2
M3
M4
P2
P21
P20
P3
M15
M9
M10
P4
M12
M11
P5
M20
M21

The Pandapas network

- Simple network with only 10 genes
- Same kinetic functions as in AGN
- Similar time scales for all genes
- Different range of values for each gene



Summary

- In silico networks provide challenging data for reverse engineering...
- They are well-suited to be gold standards
- The details of the simulated experiments may affect the possibility of having objective metrics

- In silico networks should be used for assessment of reverse engineering algorithms (but not exclusively!)

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Acknowledgements

Pedro Mendes
Diogo Camacho
Hui Cheng
Autumn Clapp
Ki m Heard
Stefan Hoops
Adaoha Ihekwaba
Aejaaz Kamal
Xing Jing Li
Ana Martins
Bharat Mehrotra
Saroj Mohapatra
Revonda Pokrzwya

Wei Sha
Paul Brazhnik
Alberto de la Fuente

Reinhard Laubenbacher
Brandy Stigler
Abdul Jarrah
Elena Dimitrova
Paola Licona

EML (Heidelberg)
Ursula Kummer
Sven Sahle
Ralph Gauges
Jürgen Pahle
Katja Wegner

Funding: