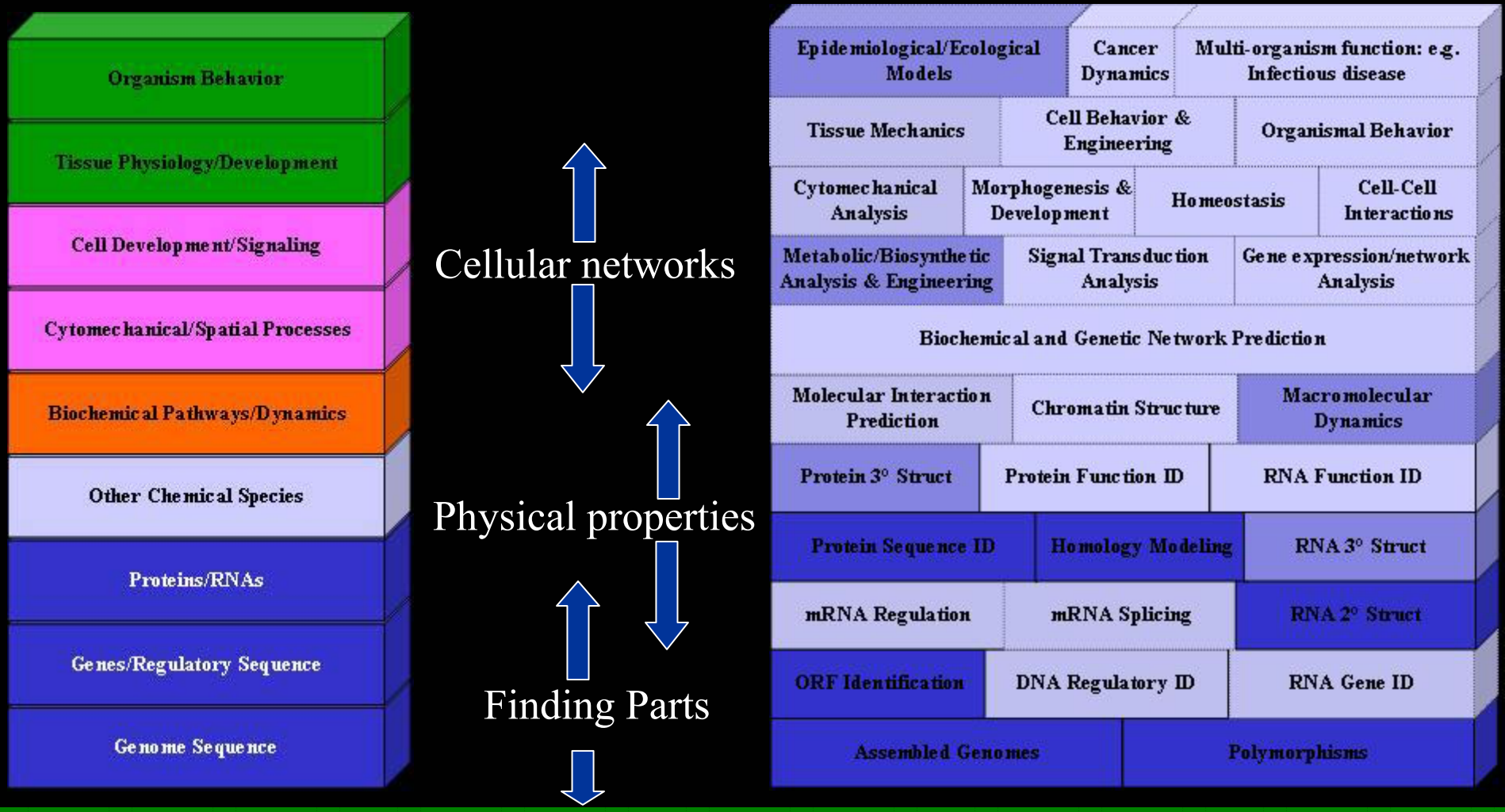
Data, Knowledge, Modules and Models

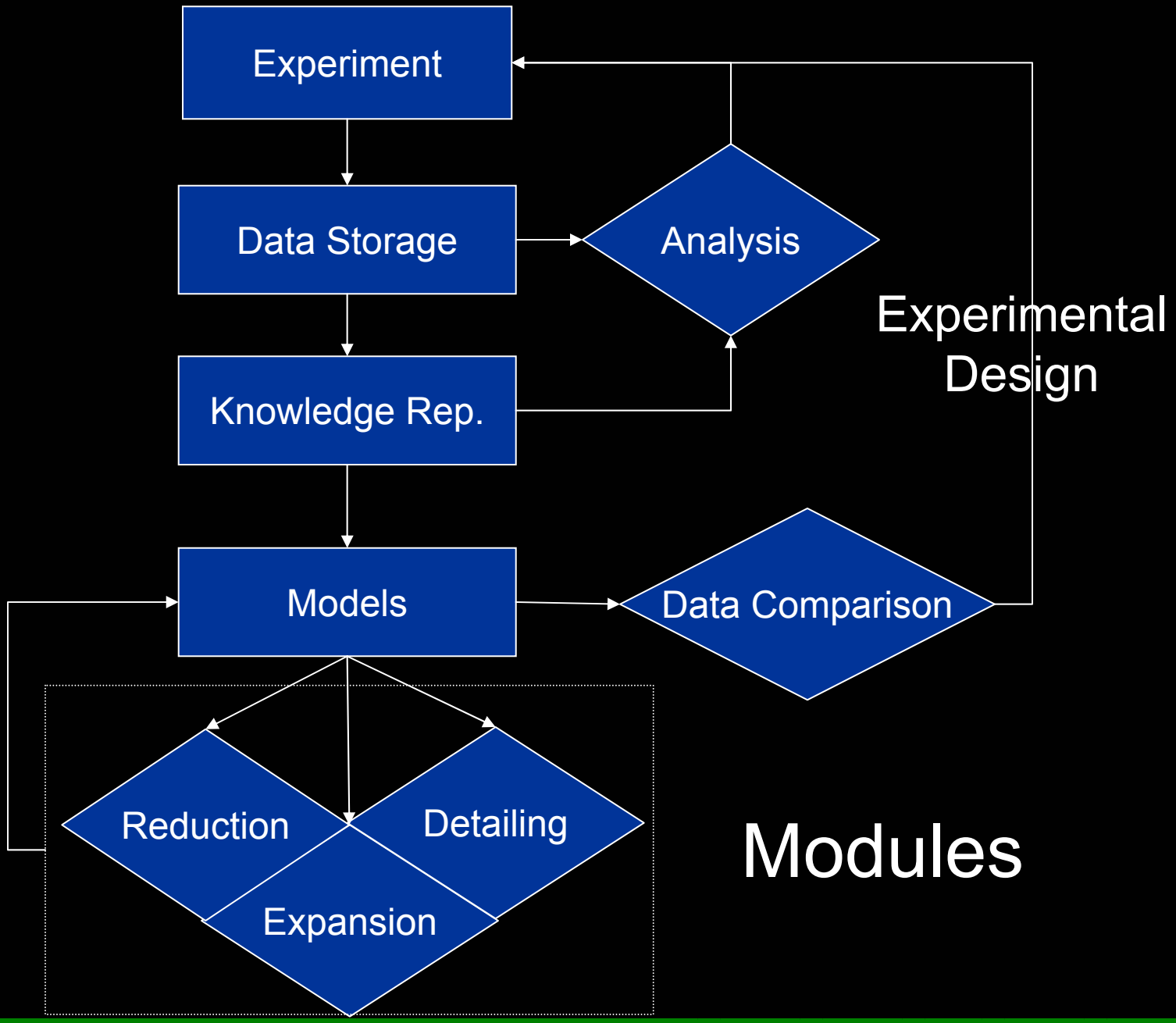
Adam Arkin
Howard Hughes Medical Institute
Departments of Bioengineering and Chemistry
University of California
Physical Biosciences Division
E.O. Lawrence Berkeley National Laboratory
Berkeley, CA 94720

http://genomics.lbl.gov

http://www.grc.uri.edu/programs/2001/bioinf.htm

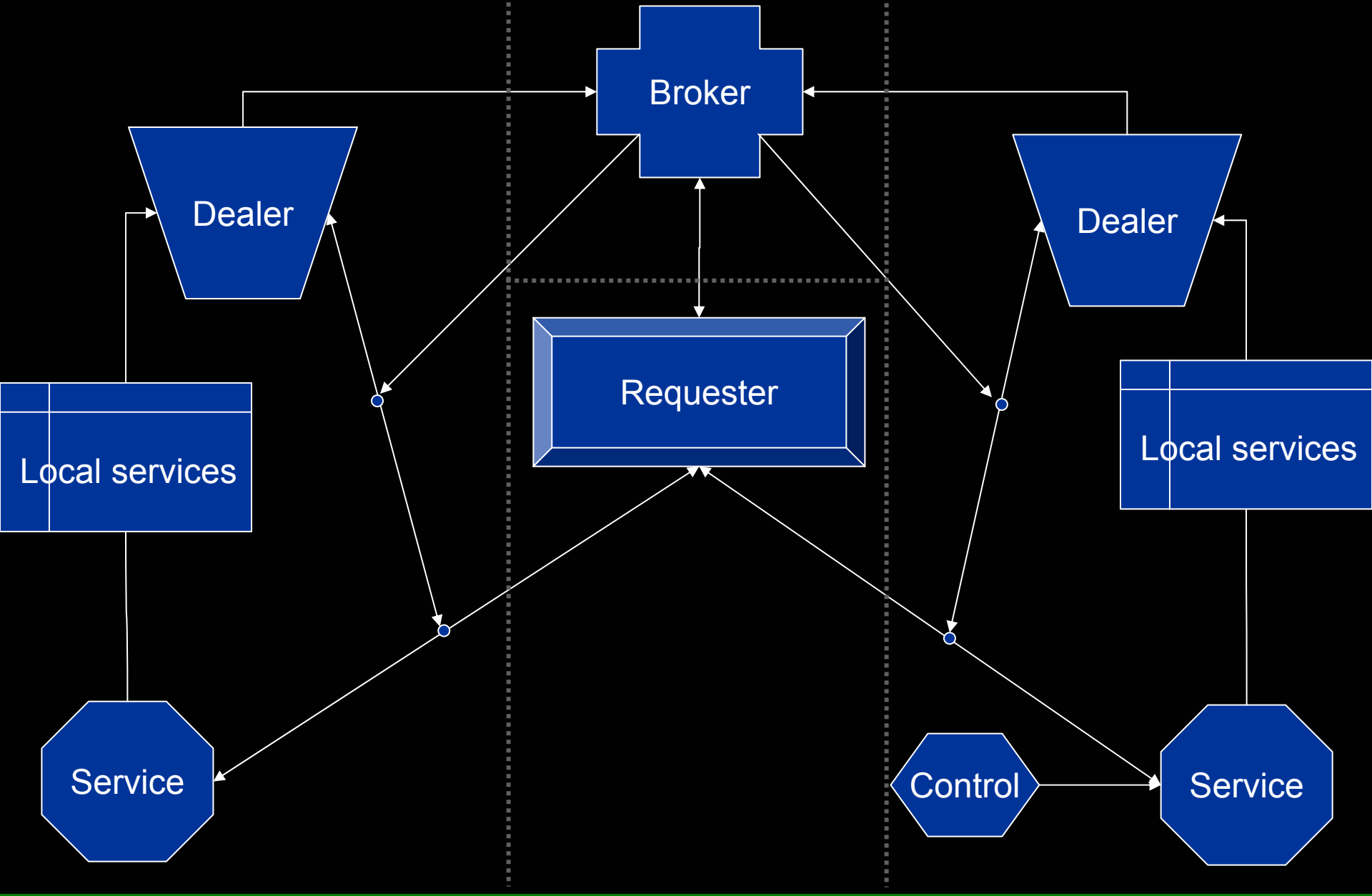
Tools for “multilevel” analysis



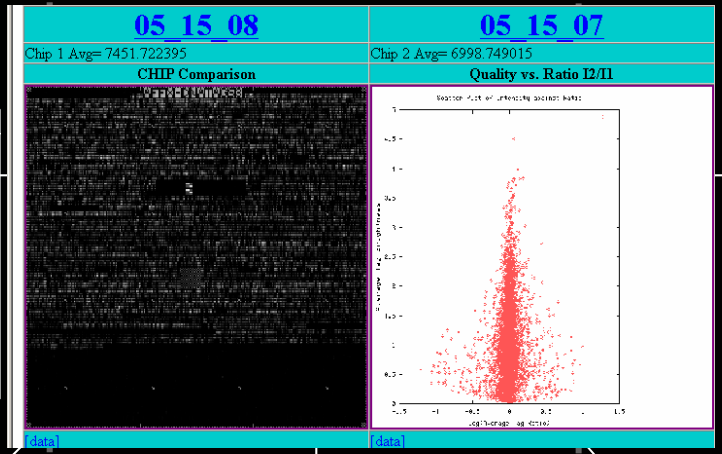

Experiment
Data Storage
Analysis
Knowledge Rep.
Experimental Design
Models
Data Comparison
Reduction
Detailing
Expansion
Modules

Building a problem solving environment

Distributed, loosely coupled, architecture



Many client services

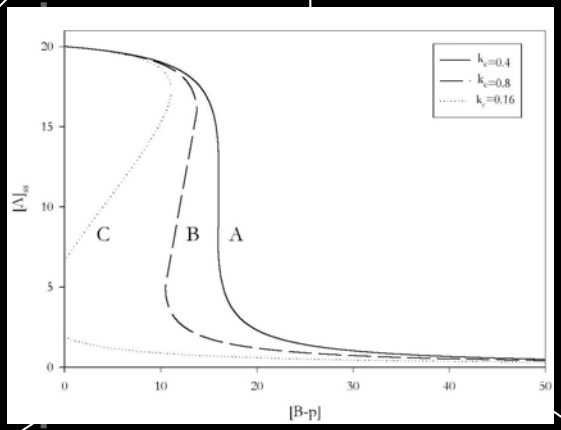


Dealer

Data Analysis

Dealer

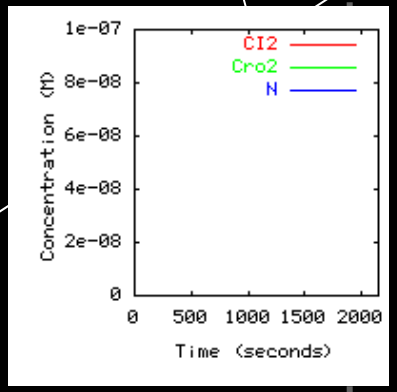
Local clients

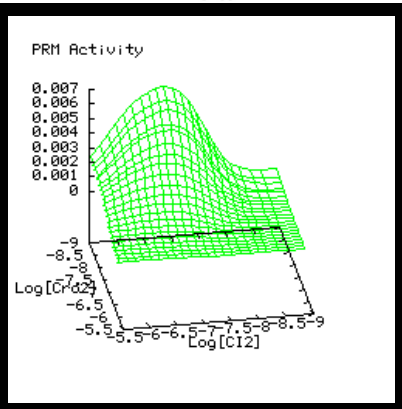


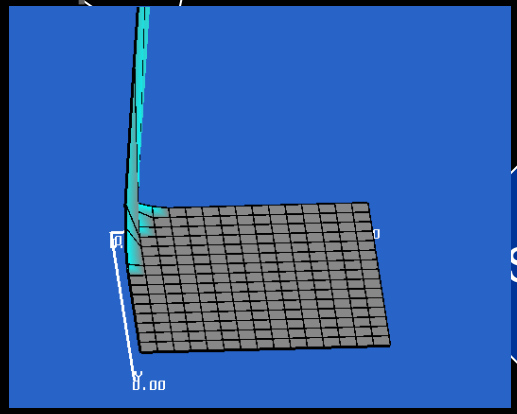
Bifurcation Analysis

Local clients

Simulation and Visualization





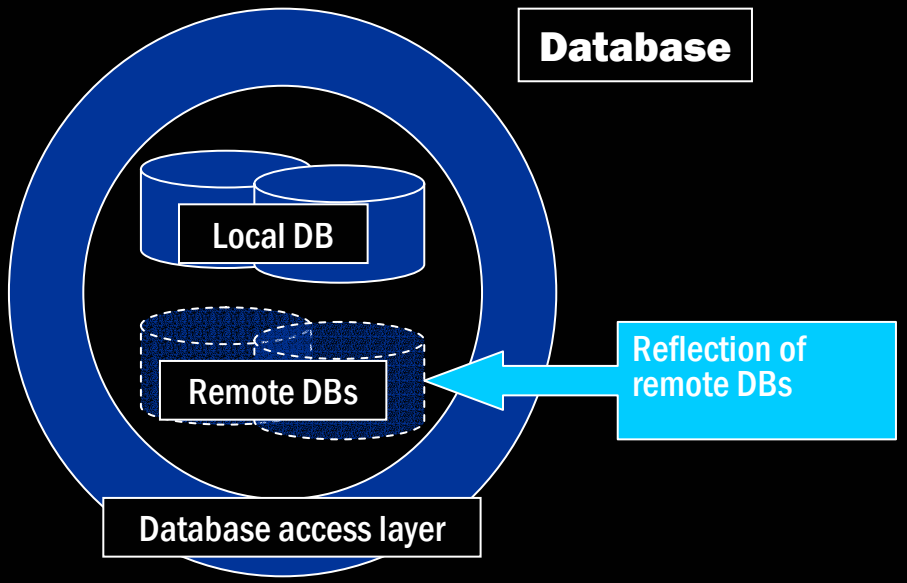


Service

Service

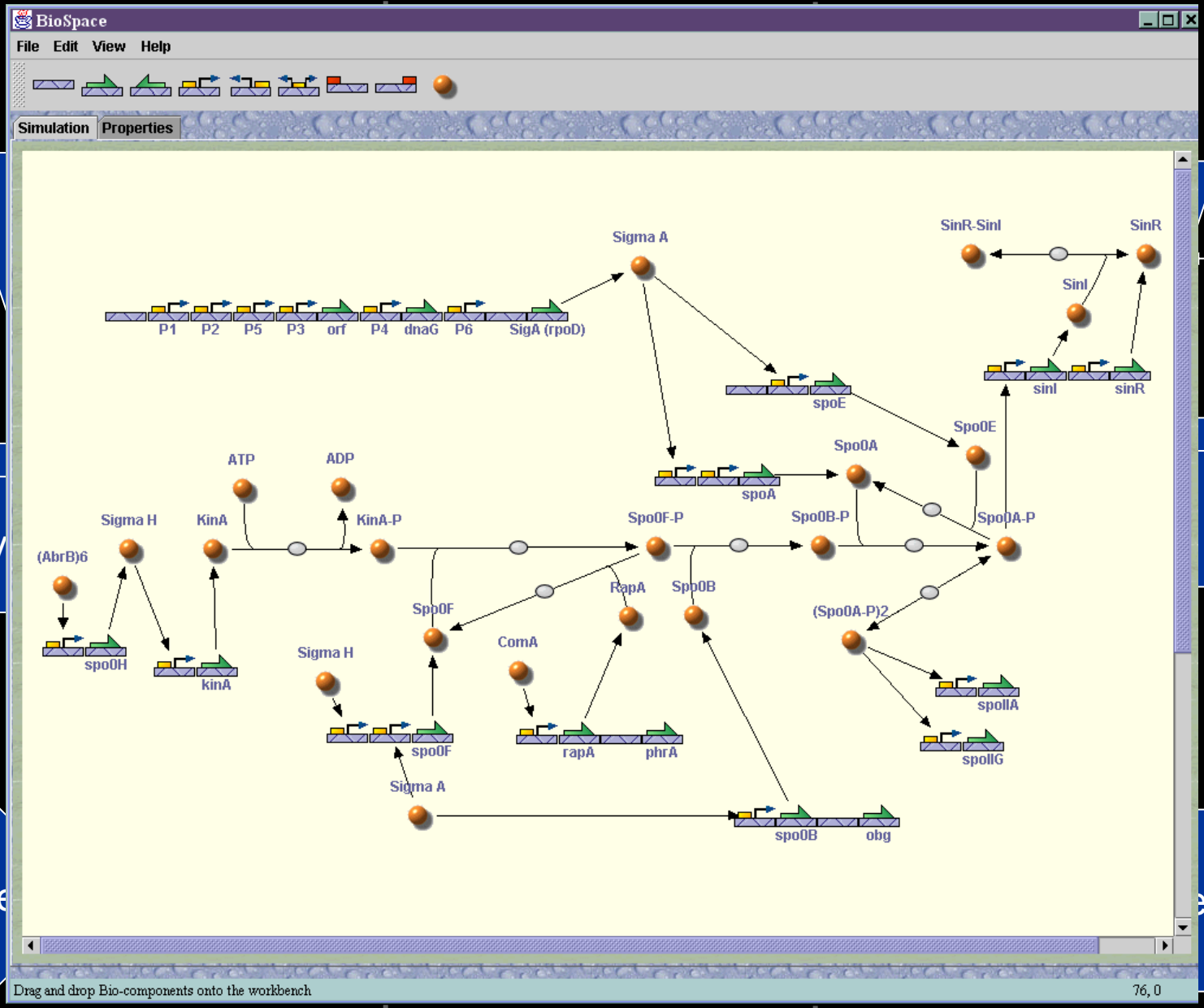
Database services

Service



Schema Compliant with
NCBI/BIND
GATC/MGED
GENBANK/PDB
AND Glue for Models
Not as nice as Shankar's

Requester is the client



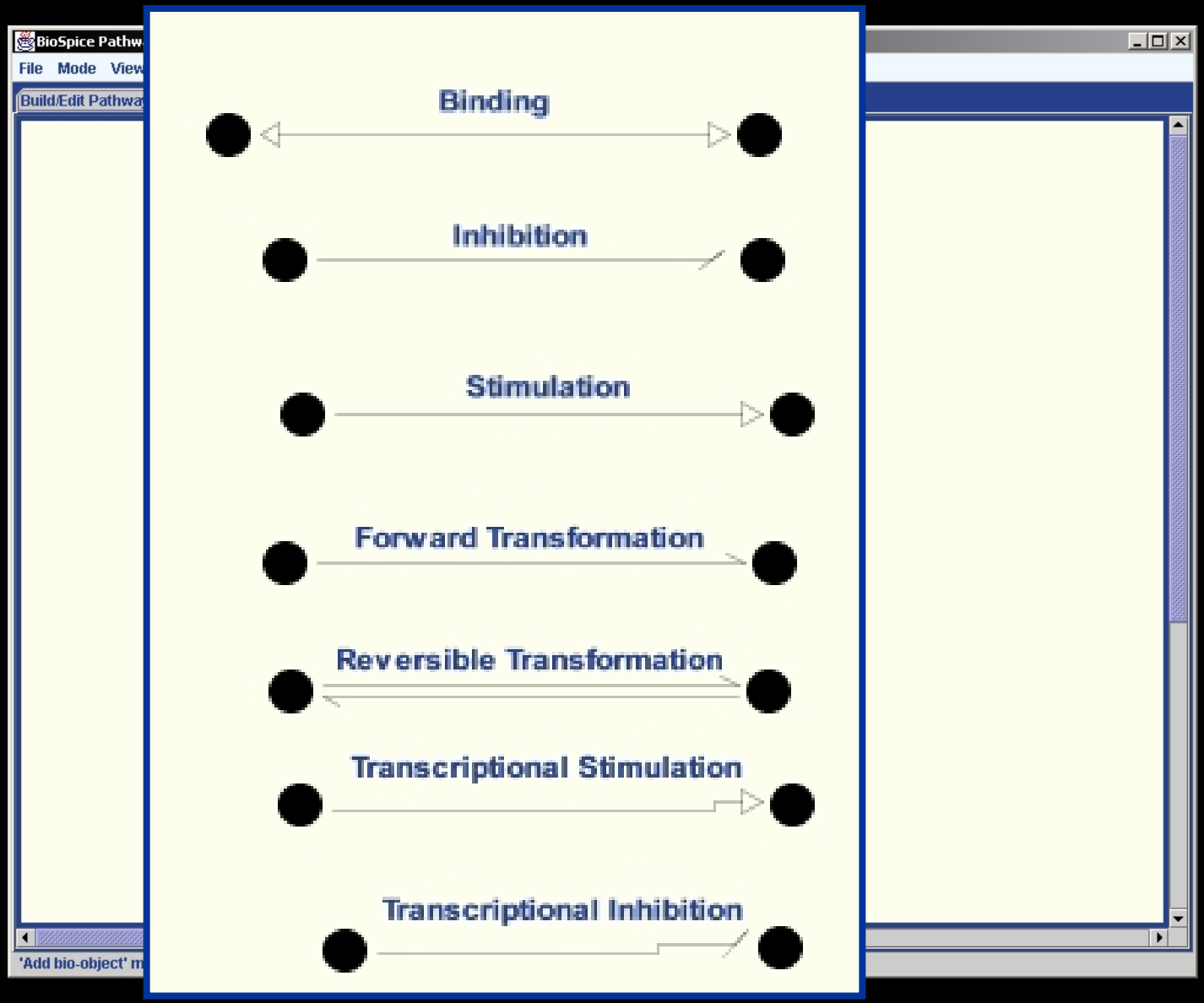
Local serv

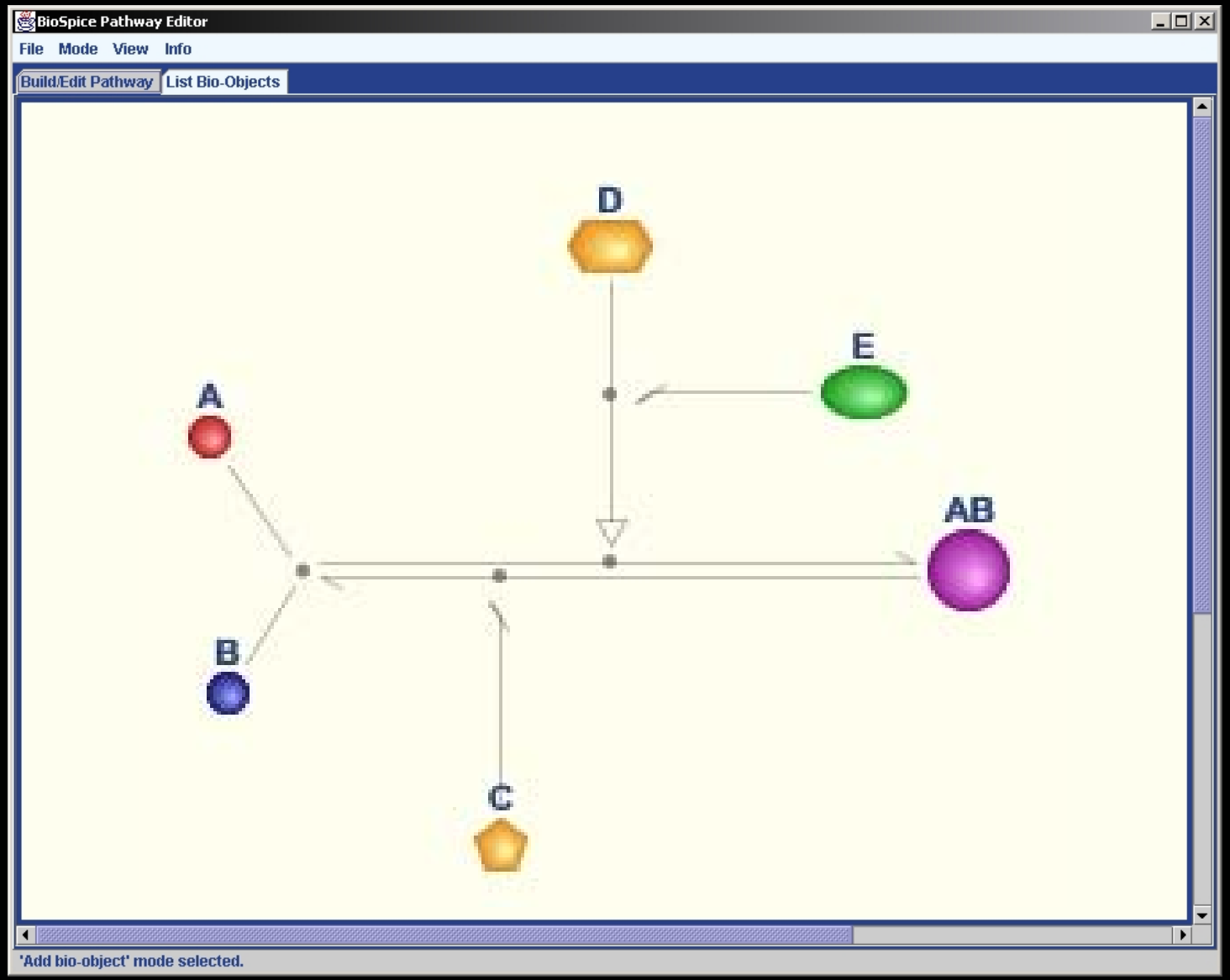
services

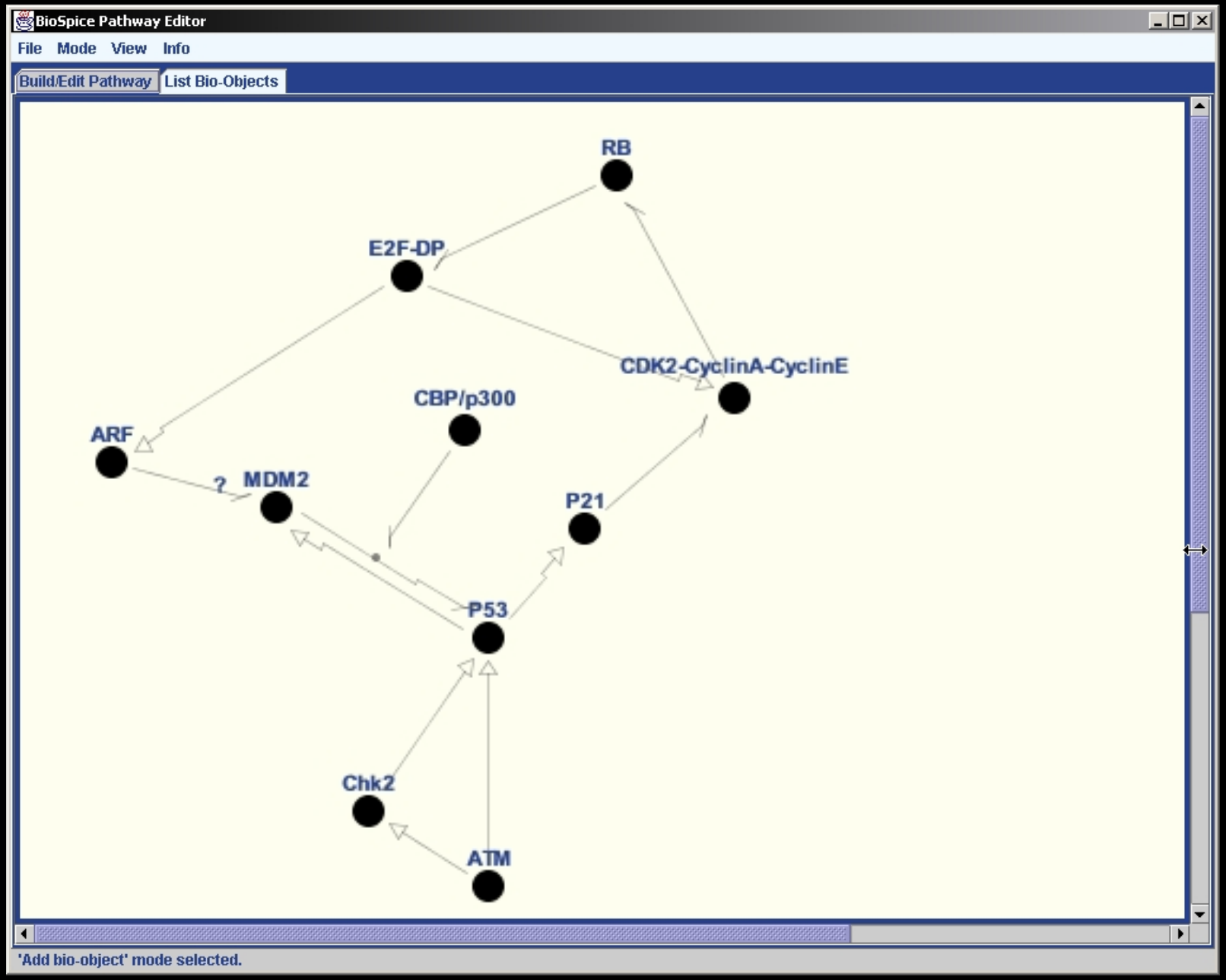
Service

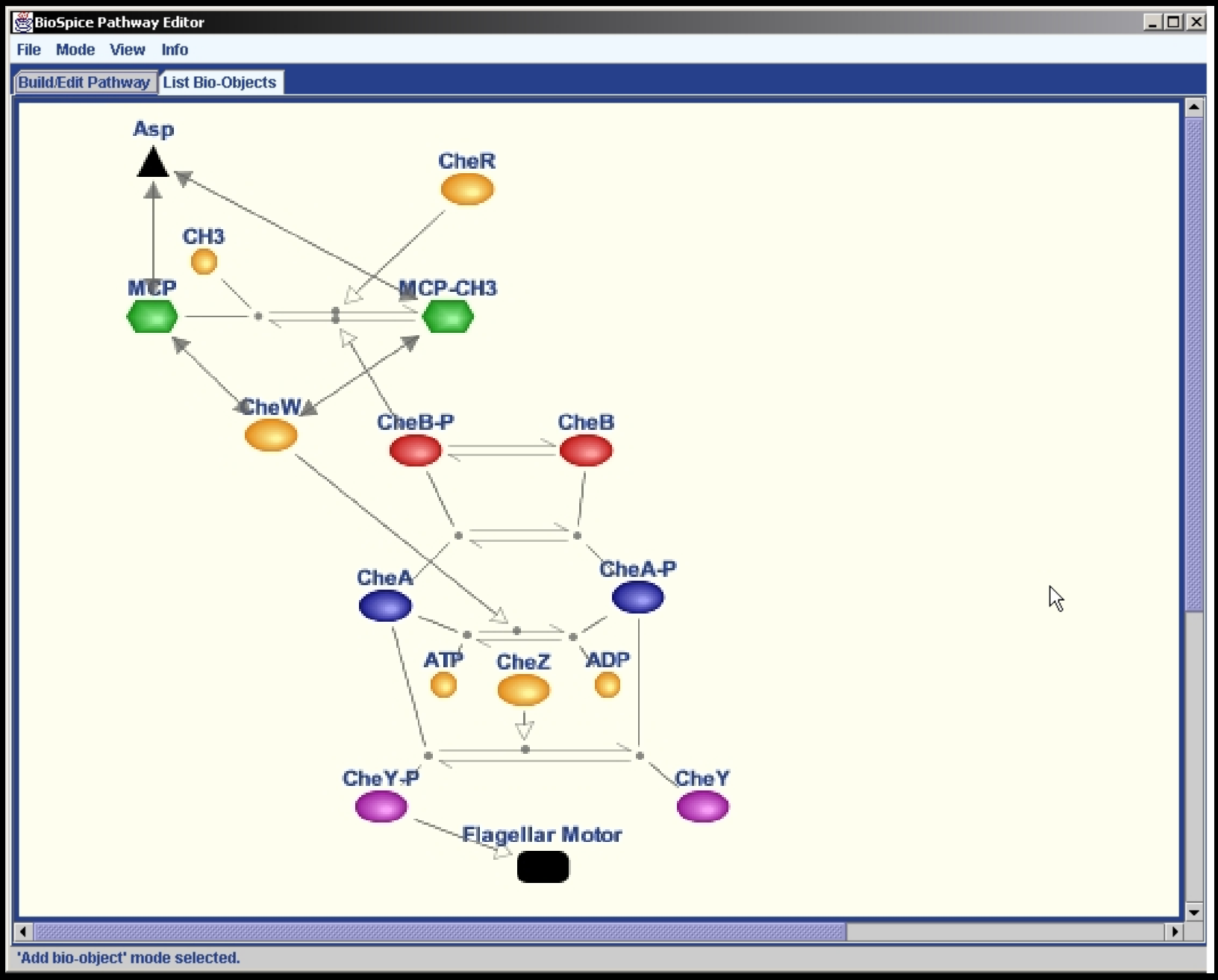
ervice

GUI must represent biological models at different levels of abstraction.




BioSpice Pathway Editor
File
Mode
View
Info
Build/Edit Pathway
List Bio-Objects
D
E
A
AB
B
C
'Add bio-object' mode selected.


BioSpice Pathway Editor
File
Mode
View
Info
Build/Edit Pathway
List Bio-Objects
RB
E2F-DP
CDK2-CyclinA-CyclinE
CBP/p300
ARF
?
MDM2
P21
P53
Chk2
ATM
'Add bio-object' mode selected.

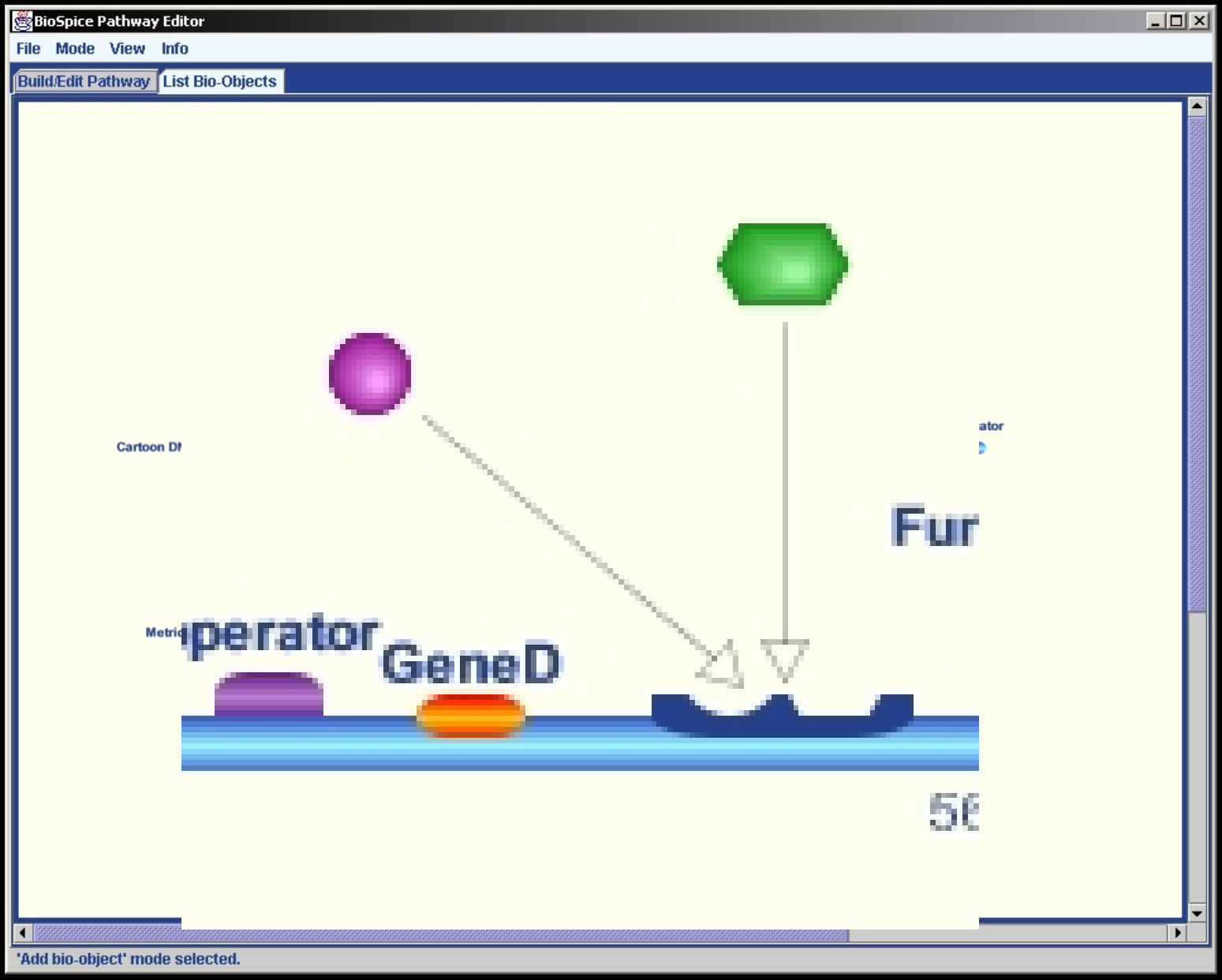

BioSpice Pathway Editor
File
Mode
View
Info
Build/Edit Pathway
List Bio-Objects
Asp
CheR
CH3
MCP
MCP-CH3
CheW
CheB-P
CheB
CheA
CheA-P
ATP
CheZ
ADP
CheY-P
CheY
Flagellar Motor
'Add bio-object' mode selected.

BioSpice Pathway Editor

File Mode View Info

Build/Edit Pathway | List Bio-Objects

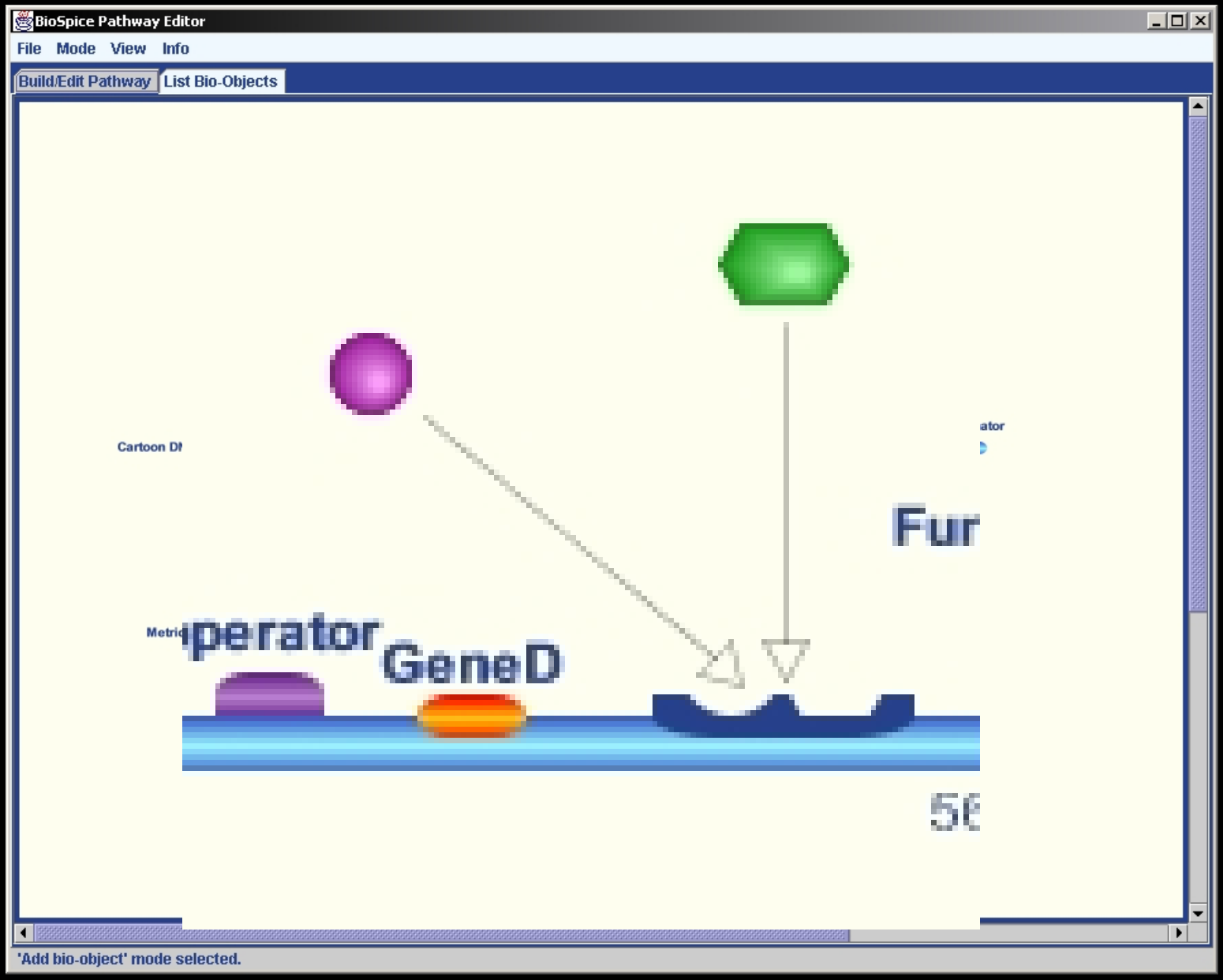
'Add bio-object' mode selected.

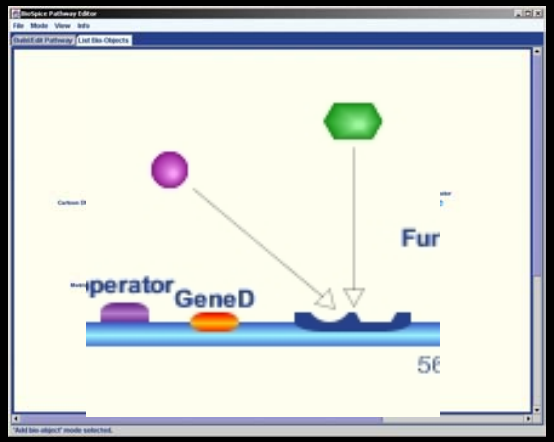

BioSpice Pathway Editor
File
Mode
View
Info
Build/Edit Pathway
List Bio-Objects
Cartoon Dl
ator
Fur
Metric
perator
GeneD
5E
'Add bio-object' mode selected.

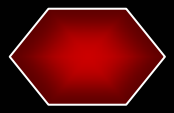
But how did we chose these icons?

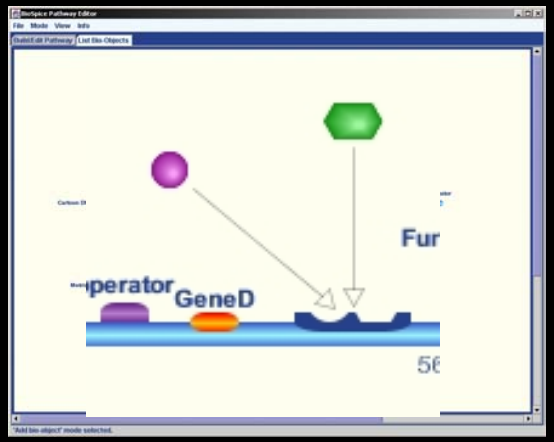
We didn’t.

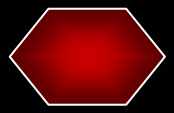
This is a big problem.

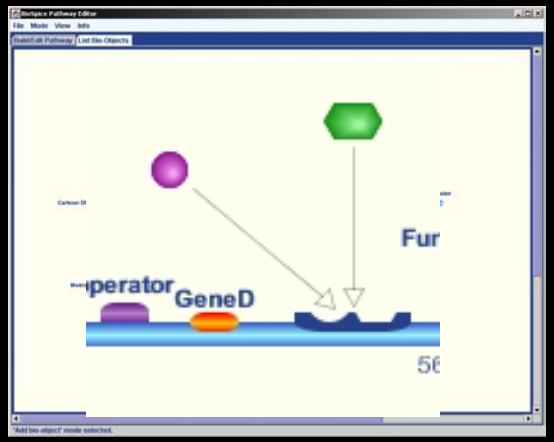

BioSpice Pathway Editor
File
Mode
View
Info
Build/Edit Pathway
List Bio-Objects
Cartoon Dr
ator
Fur
Metric
perator
GeneD
5E
'Add bio-object' mode selected.

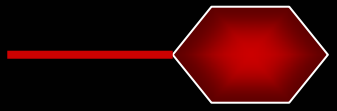

perator
GeneD
Fur




perator
GeneD
Fur




Fur
GeneD



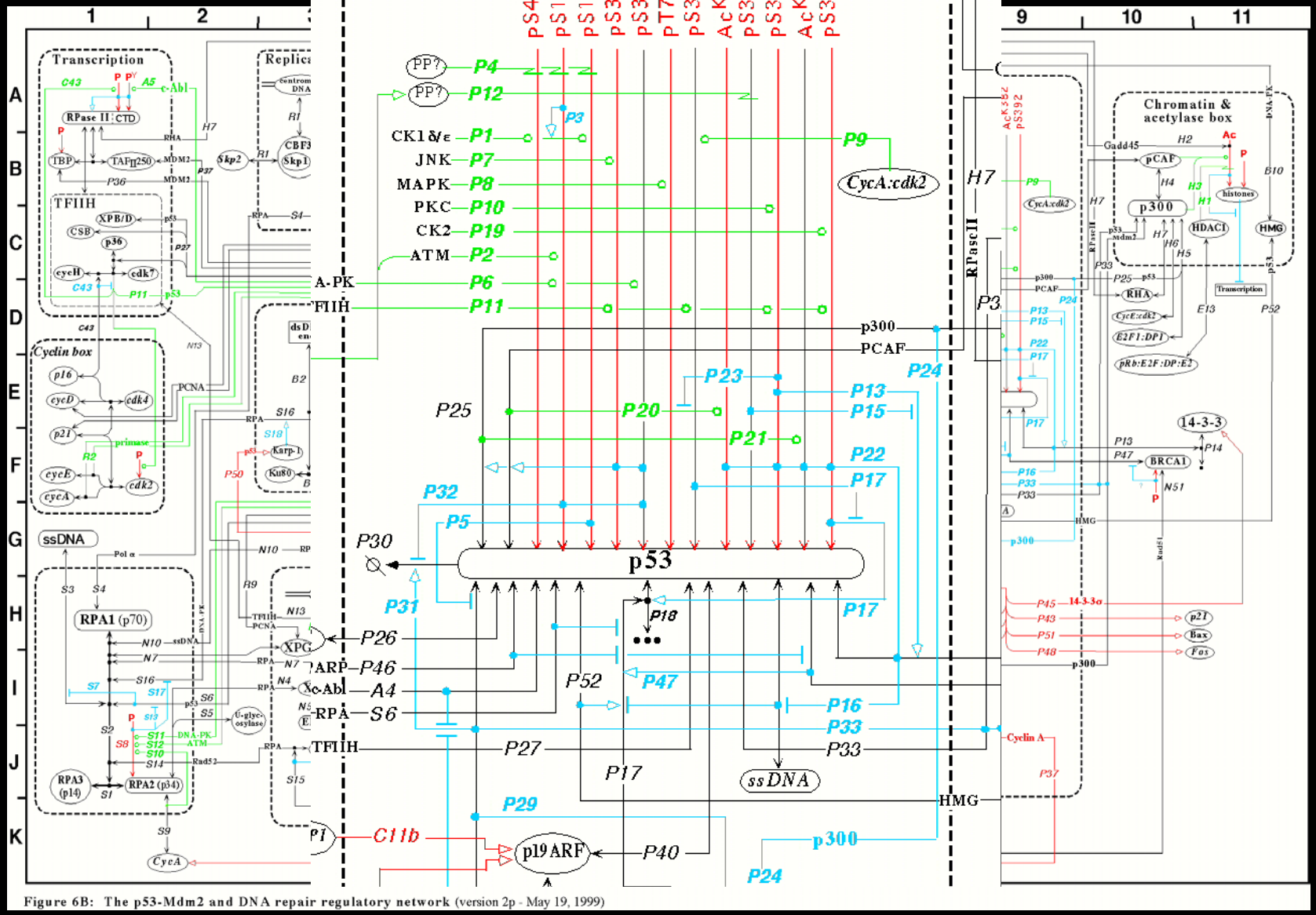
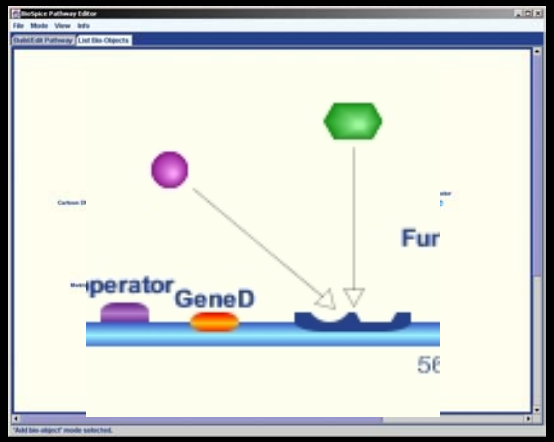


Figure 6B: The p53-Mdm2 and DNA repair regulatory network (version 2p - May 19, 1999)



So how do we represent this information?

Depends on the data

- Two/one hybrid data
- Surface Plasmon Resonance
- F.R.E.T.
- Foot Printing

Depends on the model

- Graphical
- Thermodynamic
- Kinetic

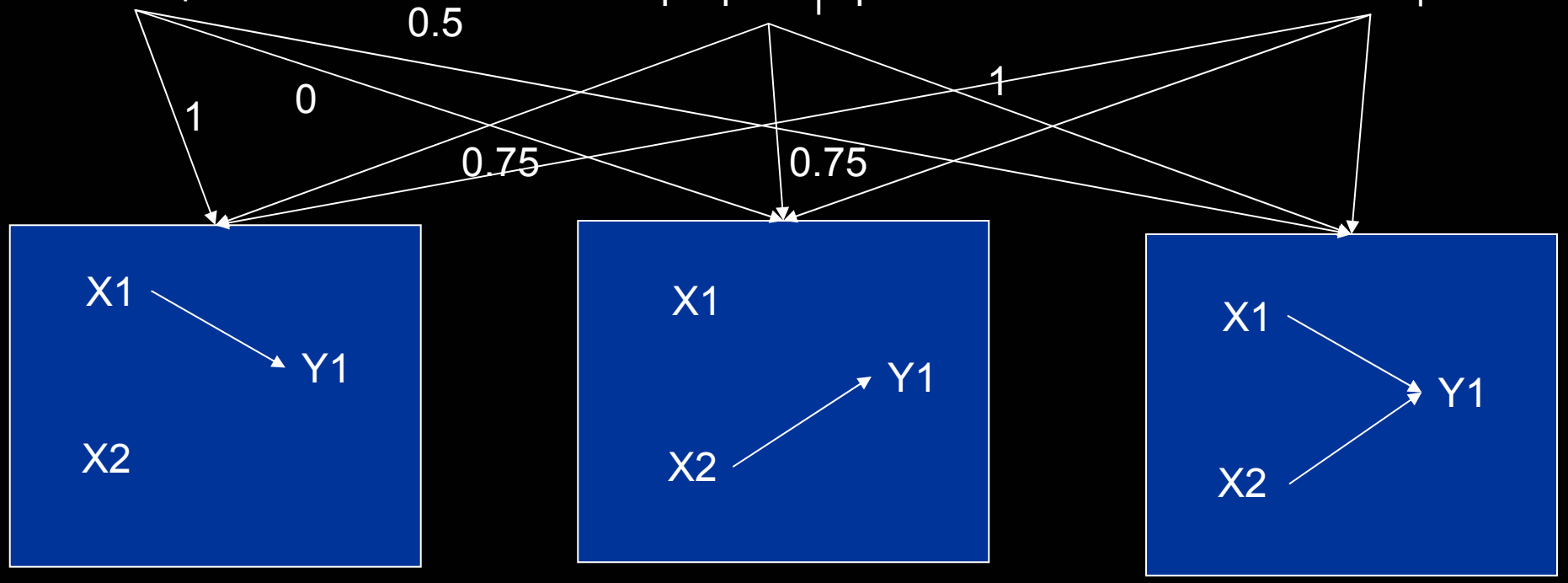
HYBRID

Graphical Models: An incredibly stupid example!

X1	X2	Y1
0	0	1
0	1	1
1	0	0
1	1	0

X1	X2	Y1
0	0	0
0	1	1
1	0	1
1	1	1

X1	X2	Y1
0	0	0
0	1	0
1	0	0
1	1	1



Obviously, our data will be more complicated

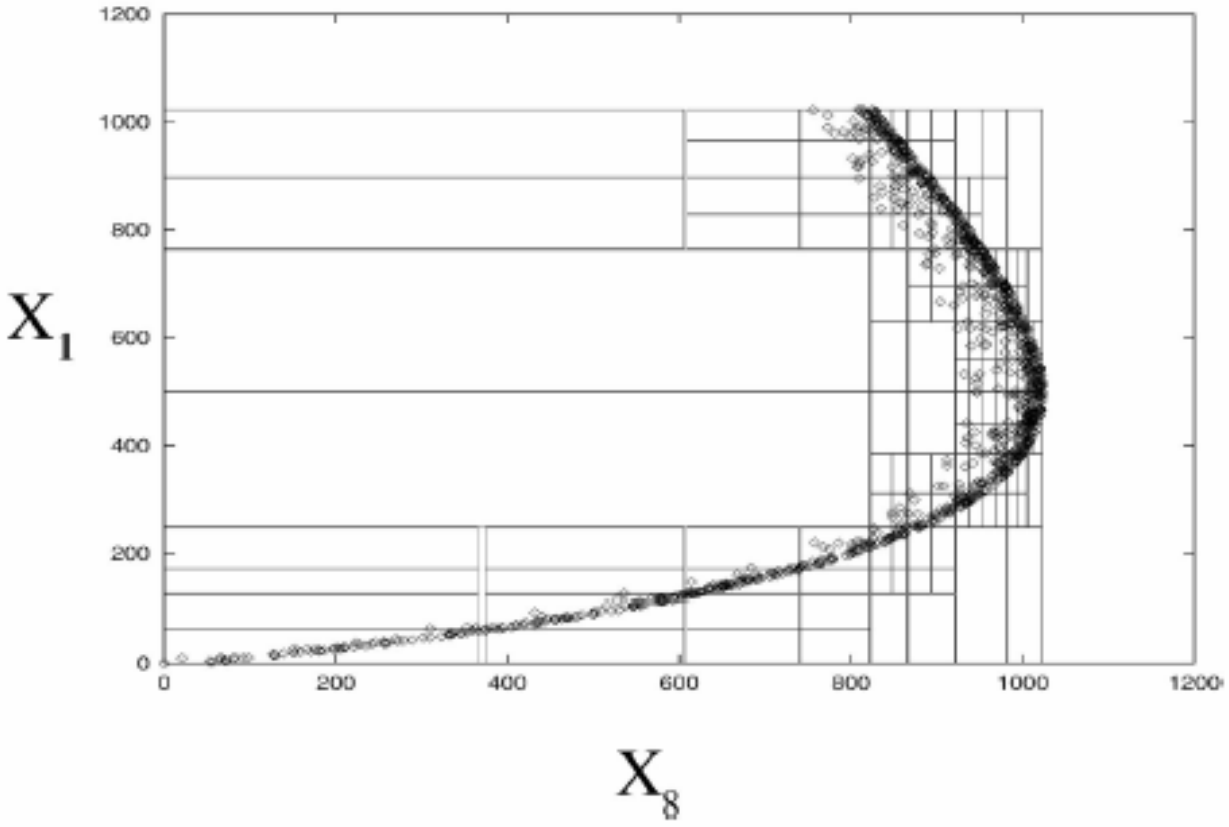


FIG. 5. The diamonds plot the values of X_1 vs X_8 obtained from the simulated time series. The rectangles are the result of a partioning algorithm, see the text. From Ref. 32.

Samoilov, Arkin, Ross, Chaos, 11(1):108-114

This gets into models and modules but…..

For graphical models data representation can be VERY simple at first.
For more complicated models we have to consider

1) What is a molecule?
   Should we represent p53 as 1 molecule with 2^{27} states?
2) What is an interaction?
   Influence?
   Direct binding?
   How do we associate different data types with it?
3) How do we relate data at different “model levels” together?

Knowledge representation for data classification and analysis

Aid to user in decision making.
Allows for data fusion.

For now:
Ontologies= Explicit specification of a conceptualization
Schema= A structure of tables in a database

Data Ontology

Analysis Ontology

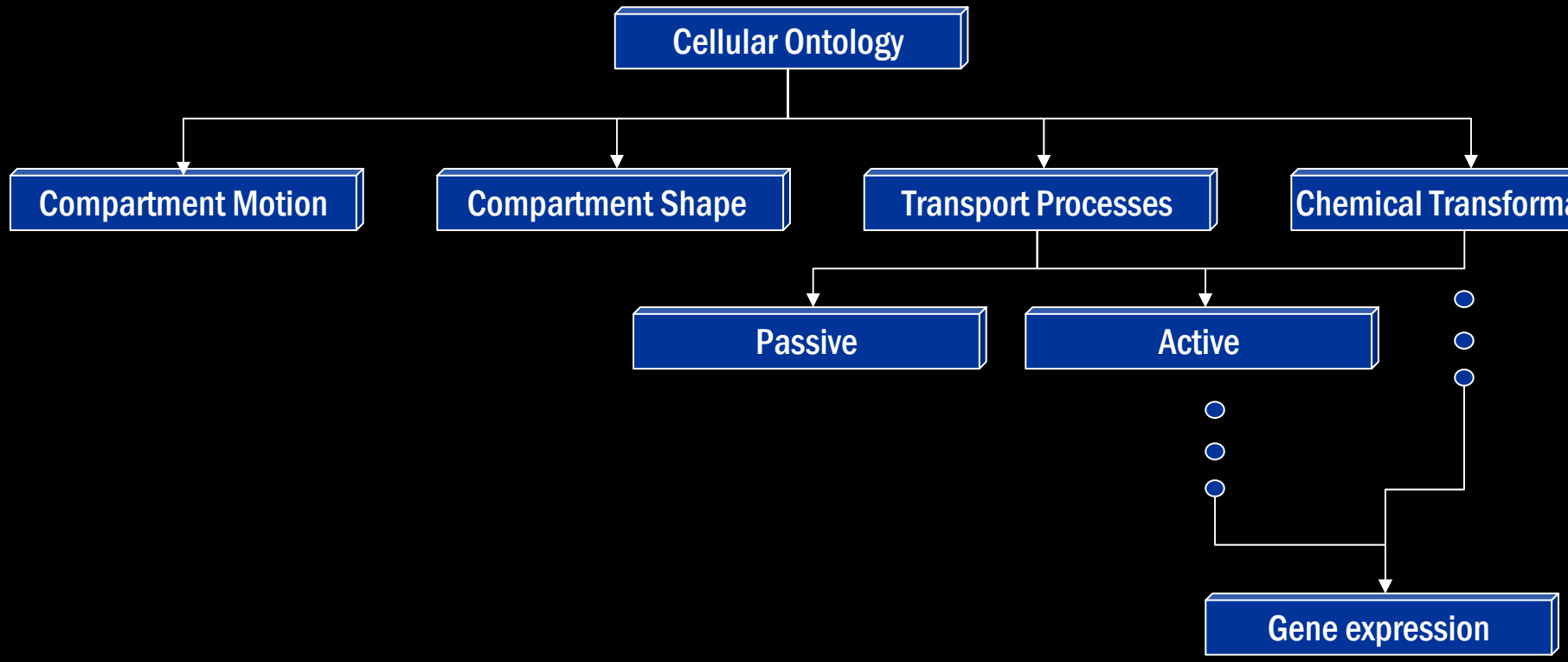
Mathematical Ontology

Differential, Algebraic, Stochastic

Cellular Ontology

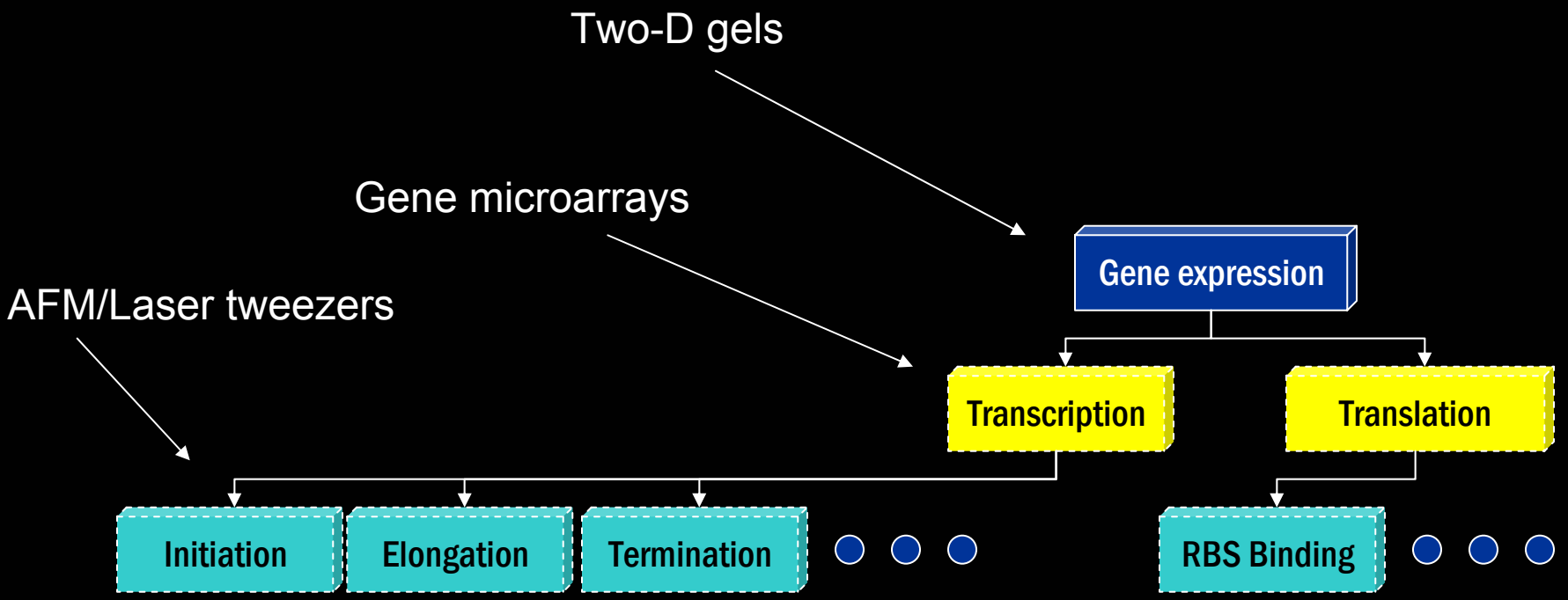
Motion, Shape Change, Transport, Transformation

Knowledge representation for data classification and analysis



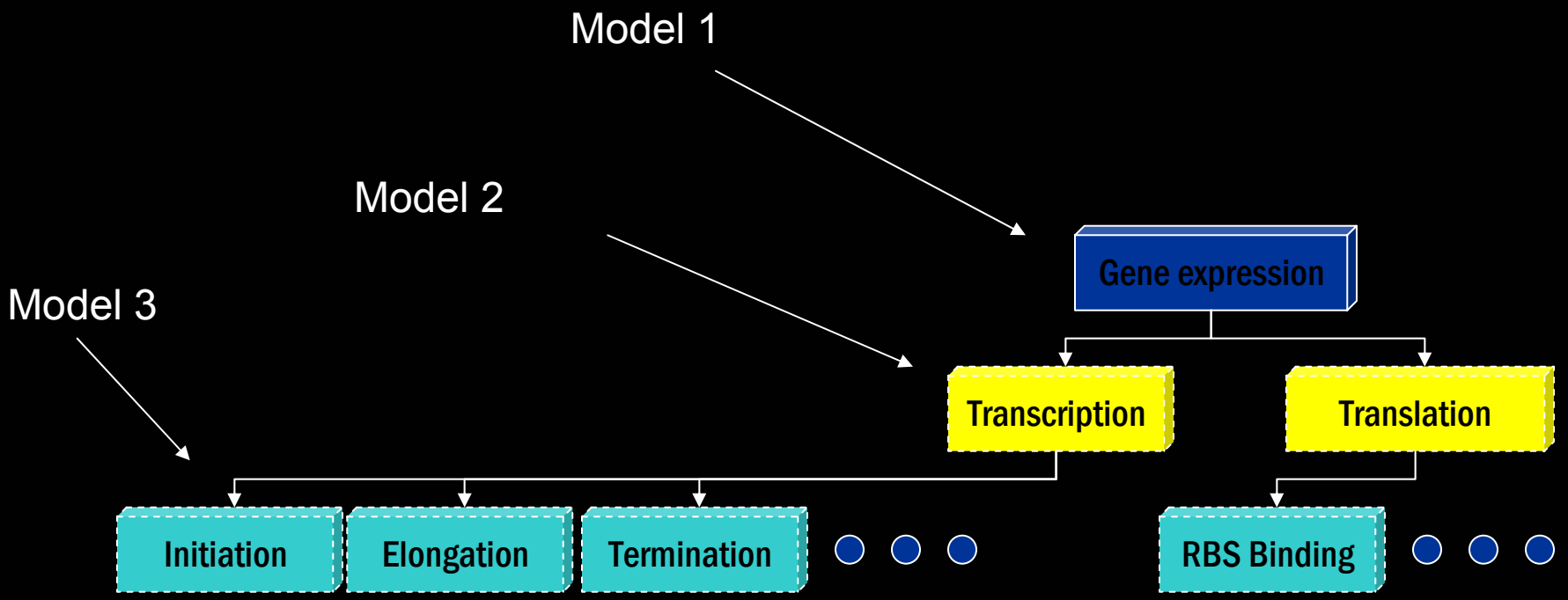
Bioontologies: http://smi-web.stanford.edu/projects/bio-ontology/

Leaves of the ontologies: Cellular

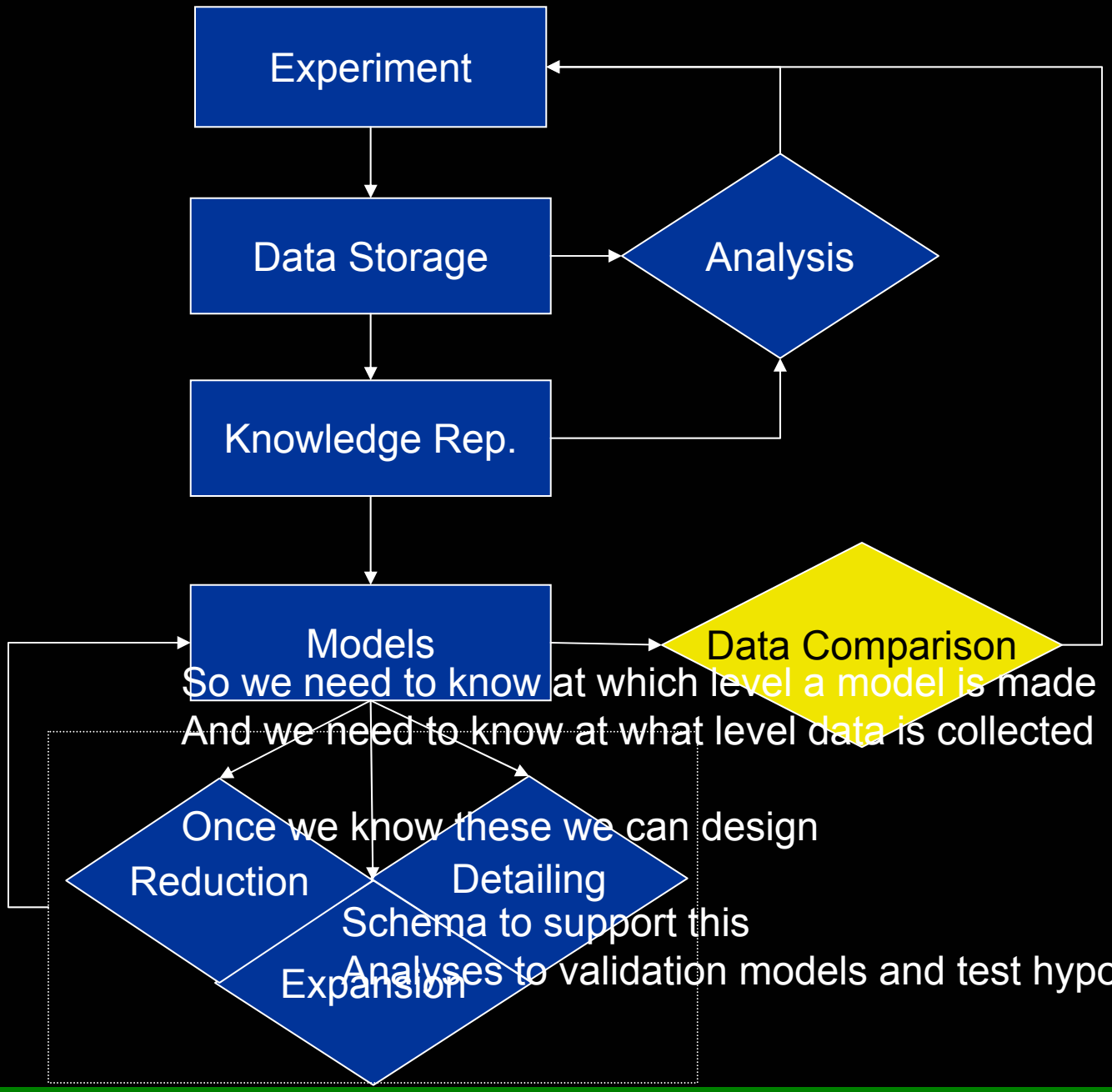


Forms a hierarchy for modeling and data

Leaves of the ontologies: Cellular



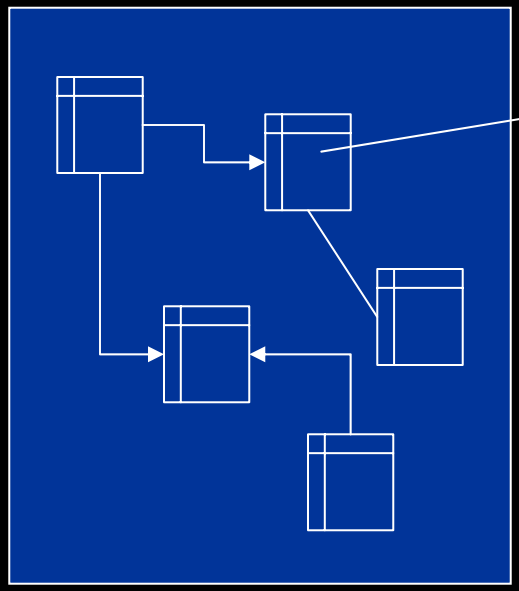
Forms a hierarchy for modeling and data



So we need to know at which level a model is made
And we need to know at what level data is collected

Once we know these we can design
Schema to support this
Analyses to validation models and test hypotheses.

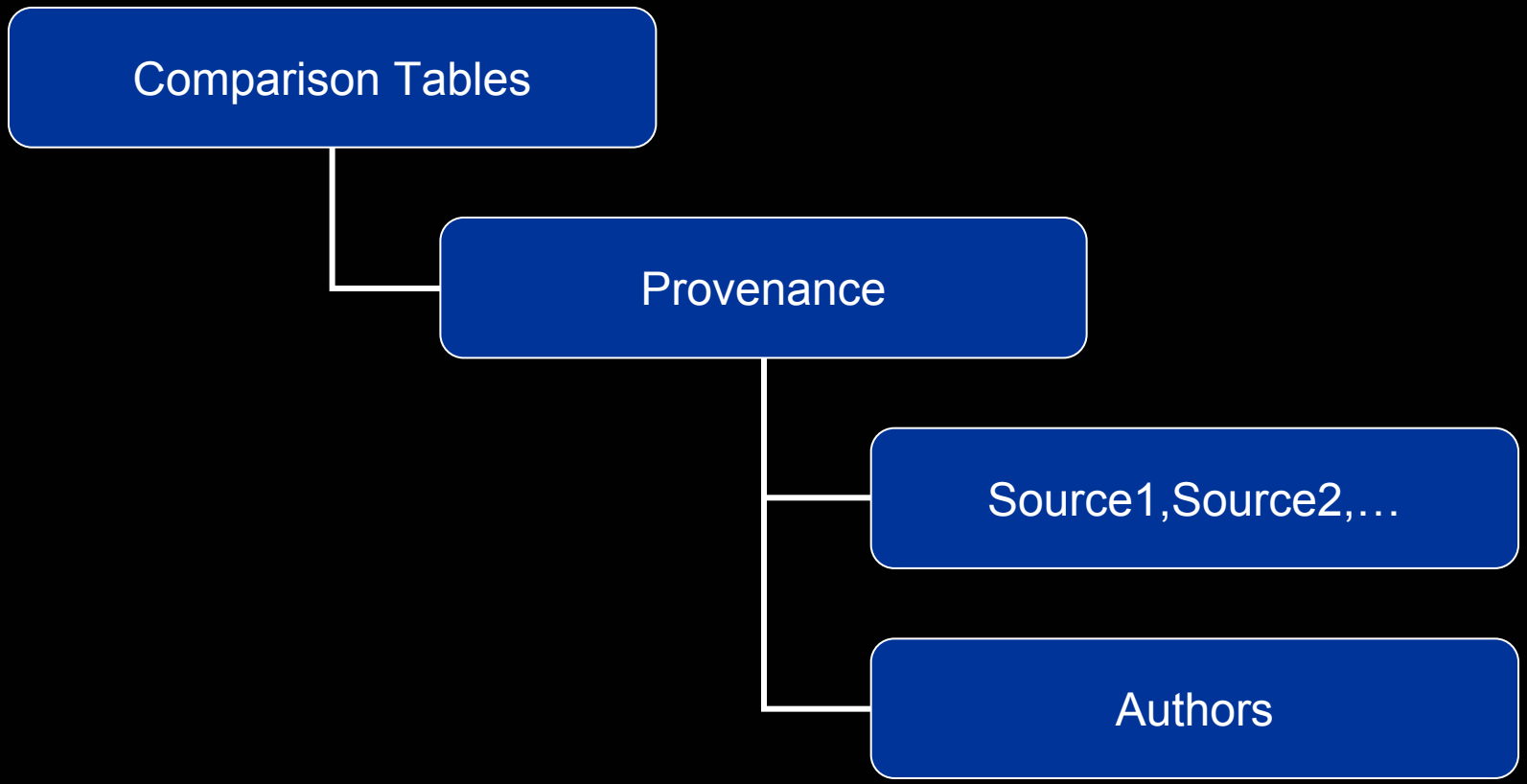
Technical fields for data comparison

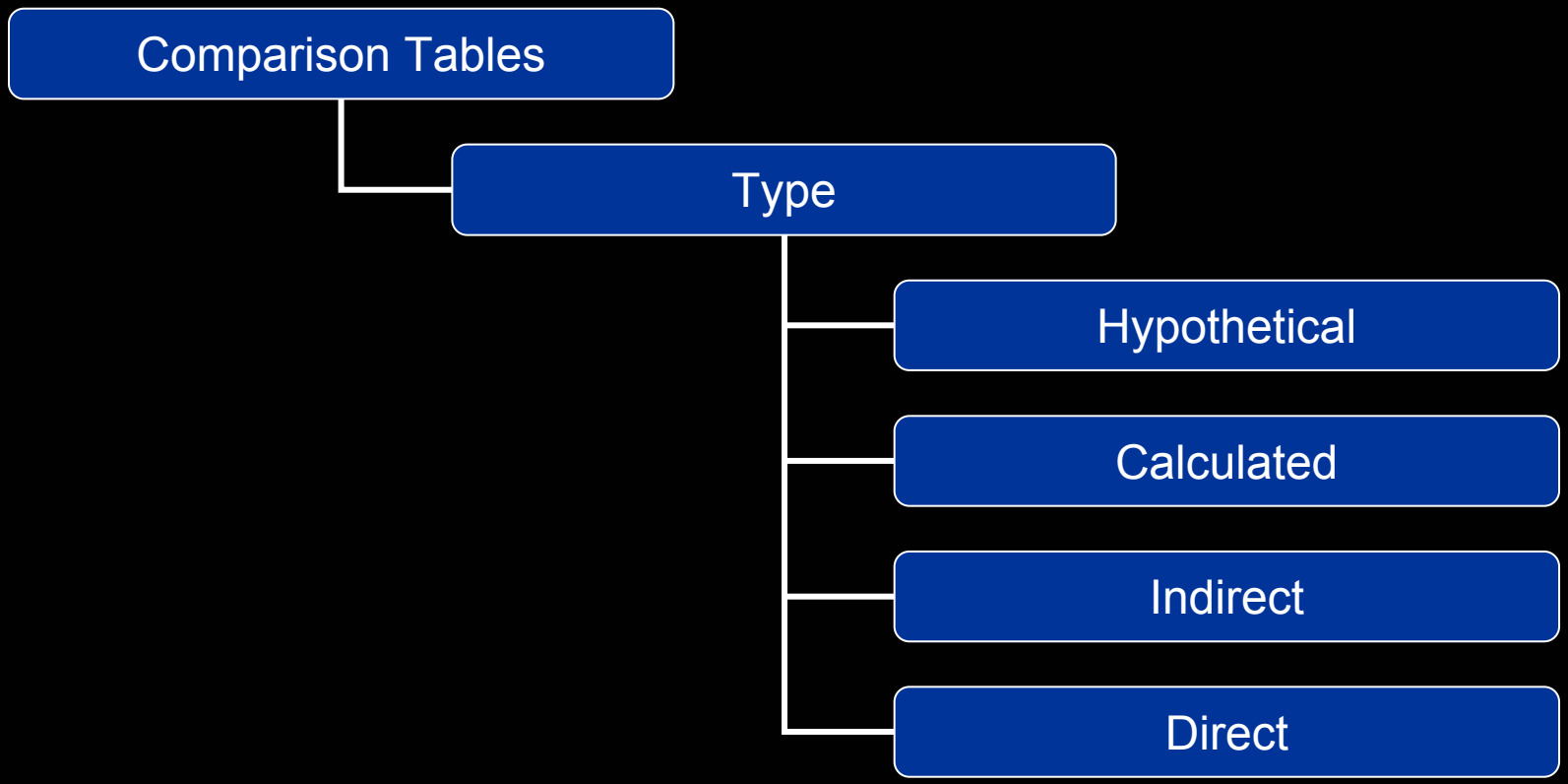


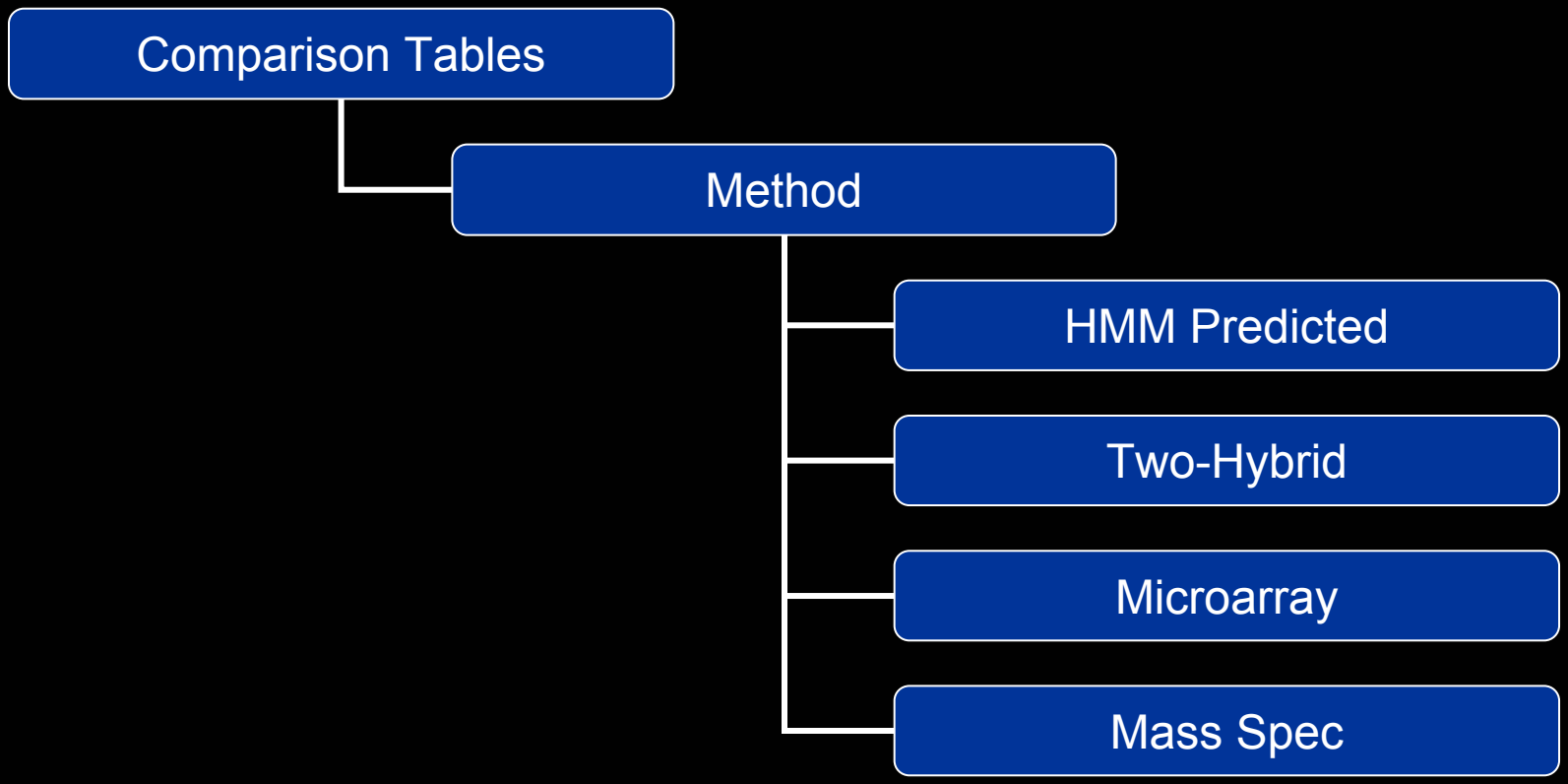
For every piece of “data” in the data base

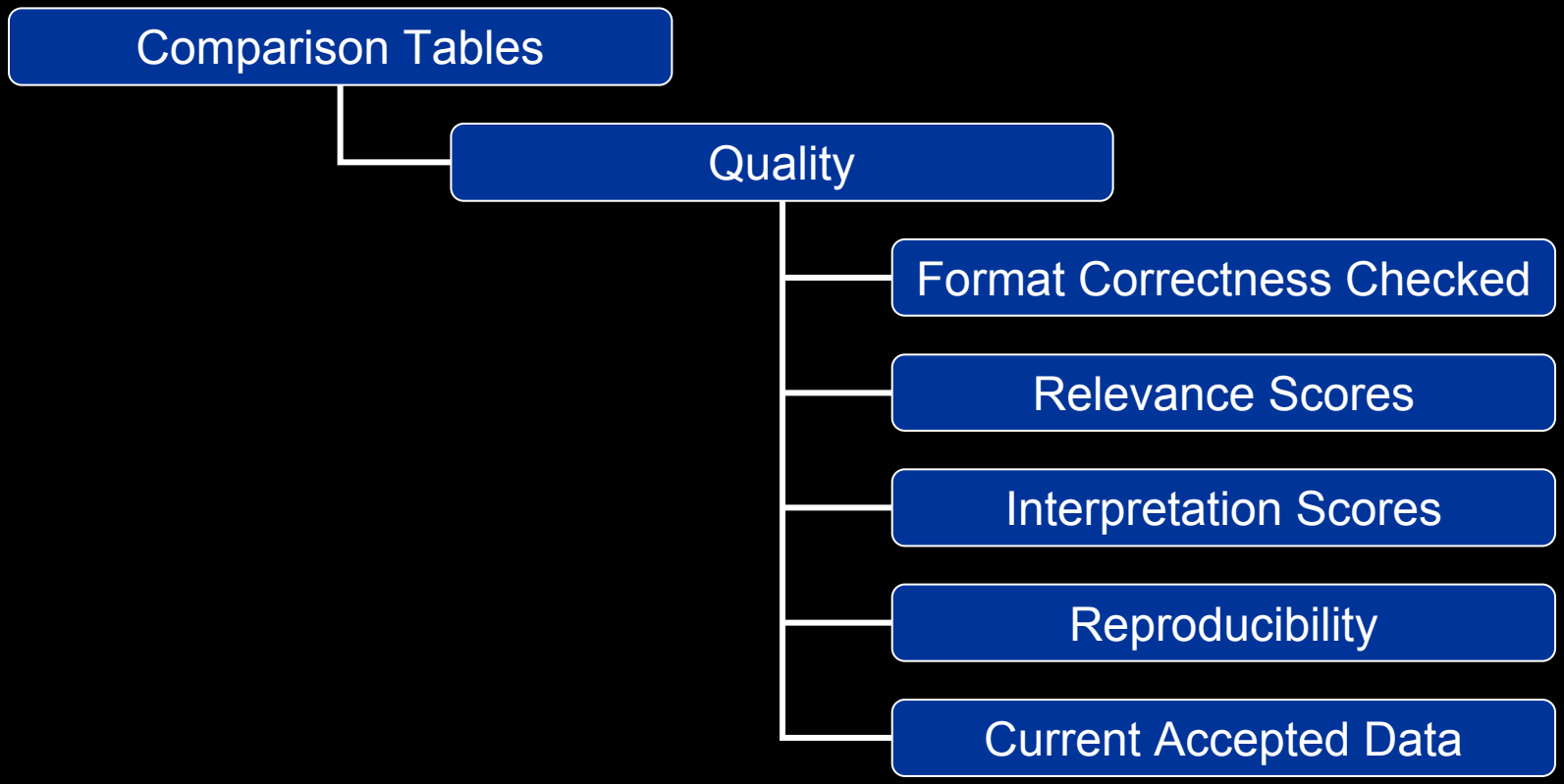
Data base schema
Assumed to contain
- Organism
- Strain
- Protocols
- Etc.

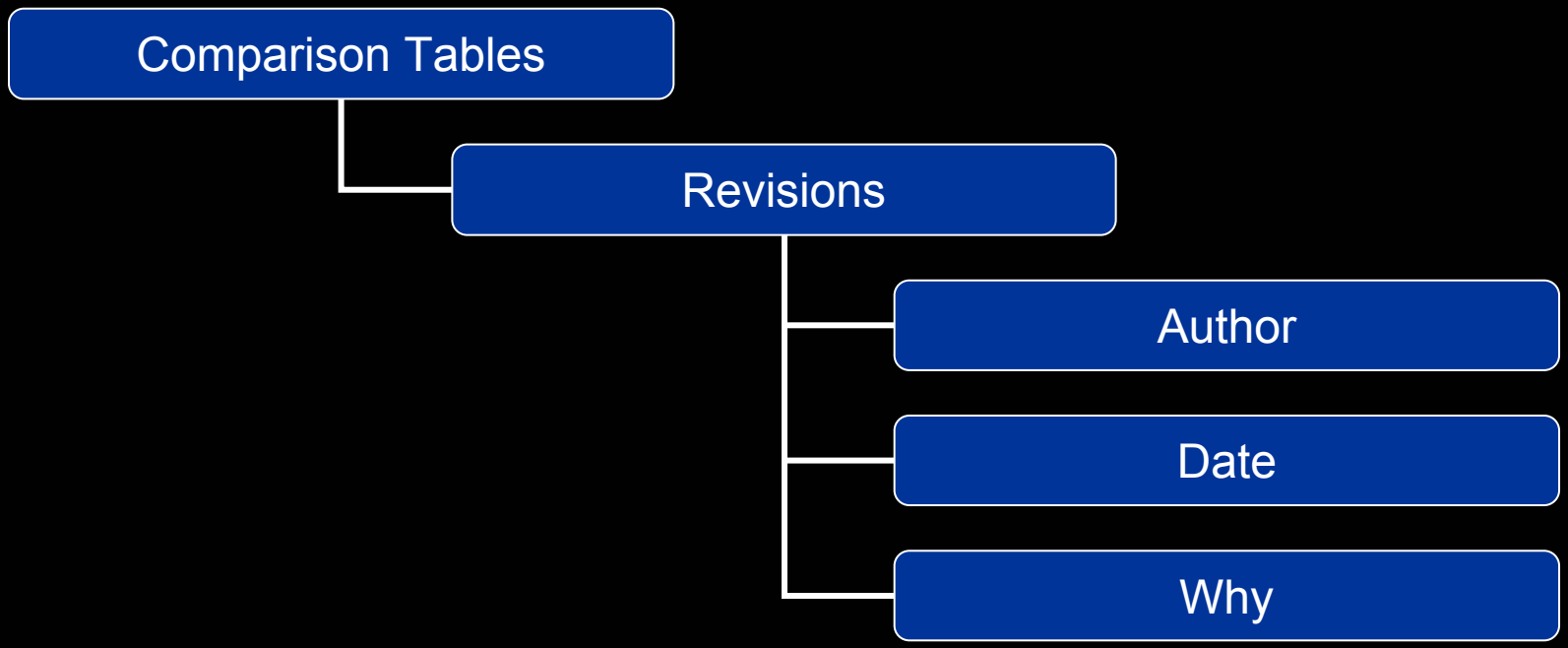
Comparison Tables
- Provenance
- Type
- Method
- Quality
- Revision
- Alternate

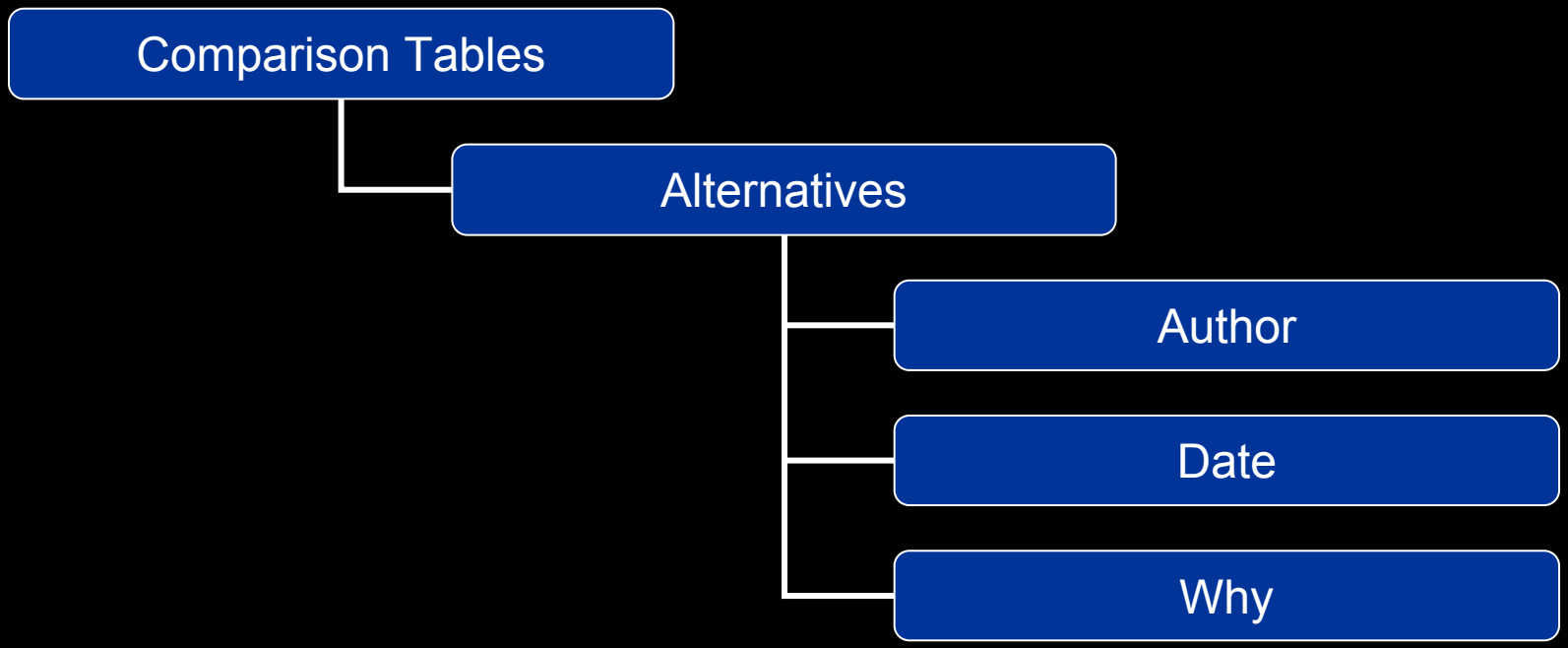

Comparison Tables
Provenance
Source1,Source2,…
Authors


Comparison Tables
Type
Hypothetical
Calculated
Indirect
Direct


Comparison Tables
Method
HMM Predicted
Two-Hybrid
Microarray
Mass Spec

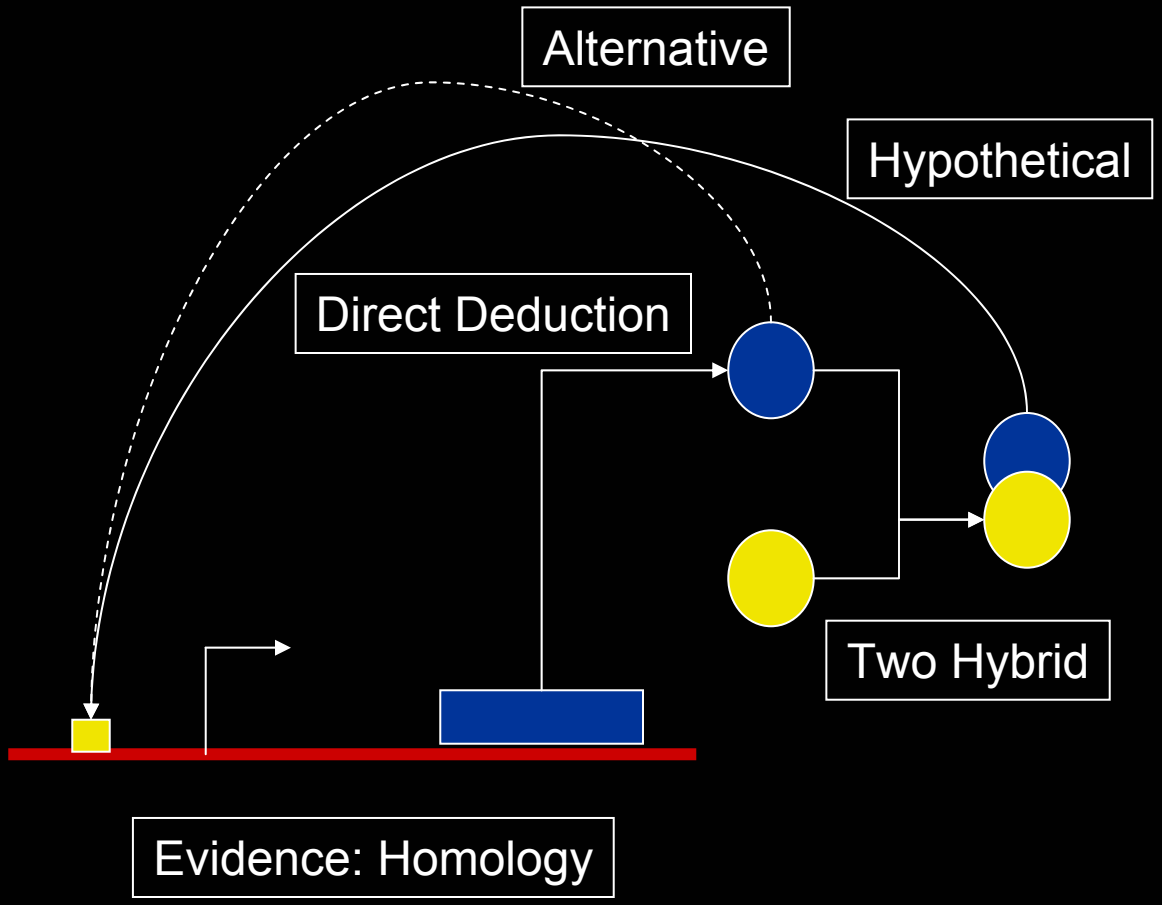

Comparison Tables
Quality
Format Correctness Checked
Relevance Scores
Interpretation Scores
Reproducibility
Current Accepted Data


Comparison Tables
Revisions
Author
Date
Why

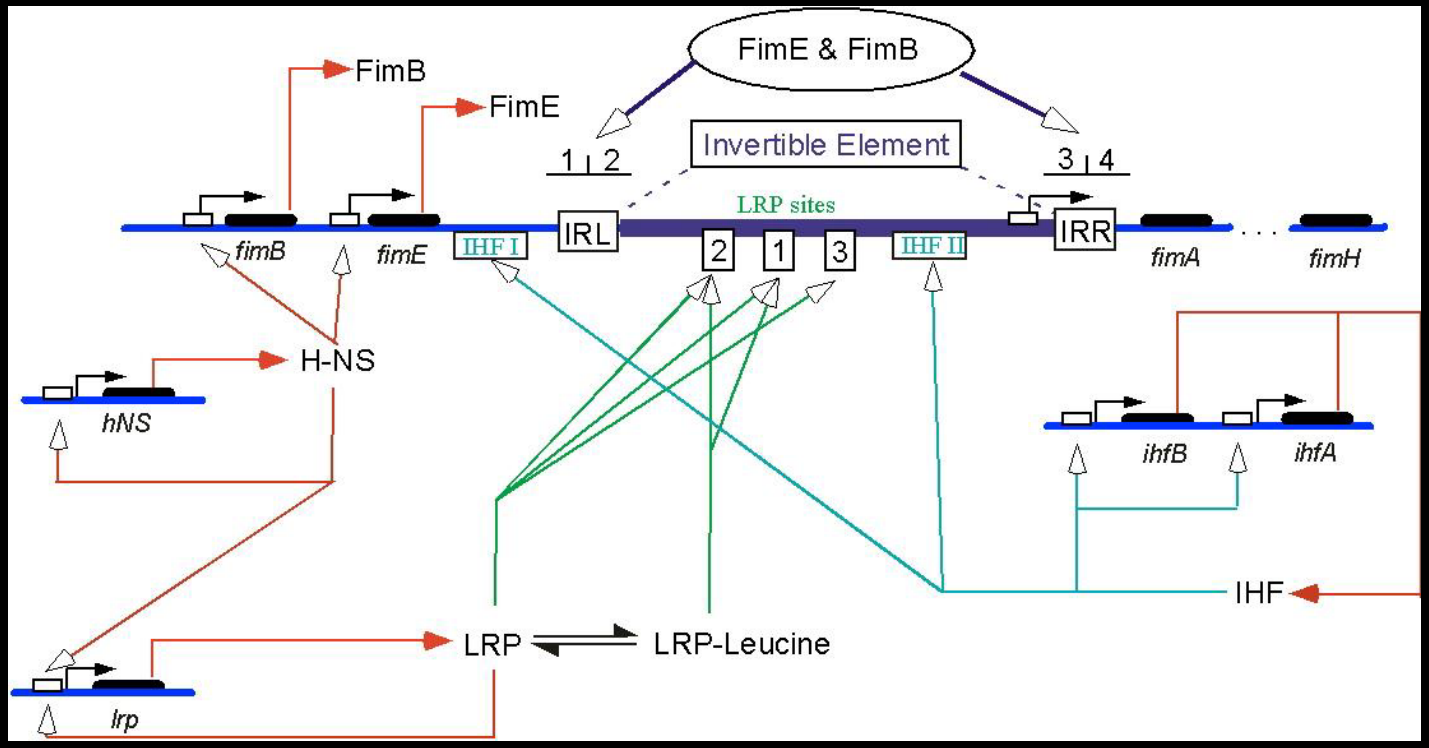

Comparison Tables
Alternatives
Author
Date
Why

Alternatives are especially important for model data bases.

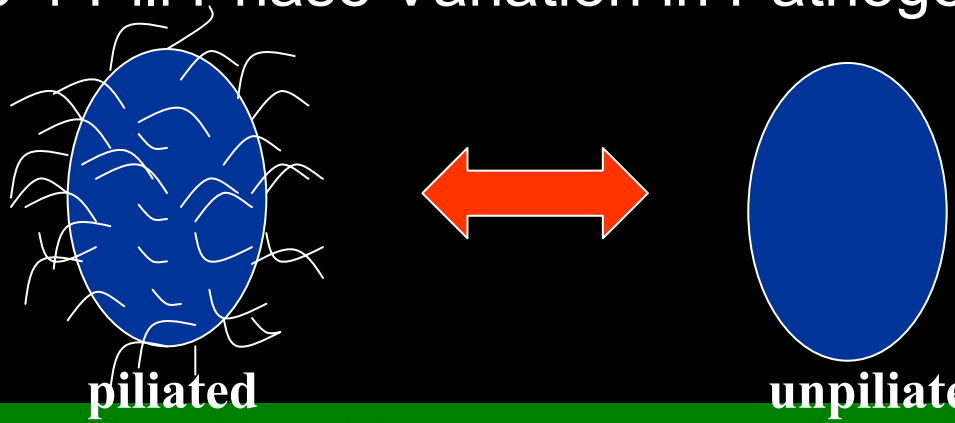
A model is a collection of data of a particular sort and hypotheses.



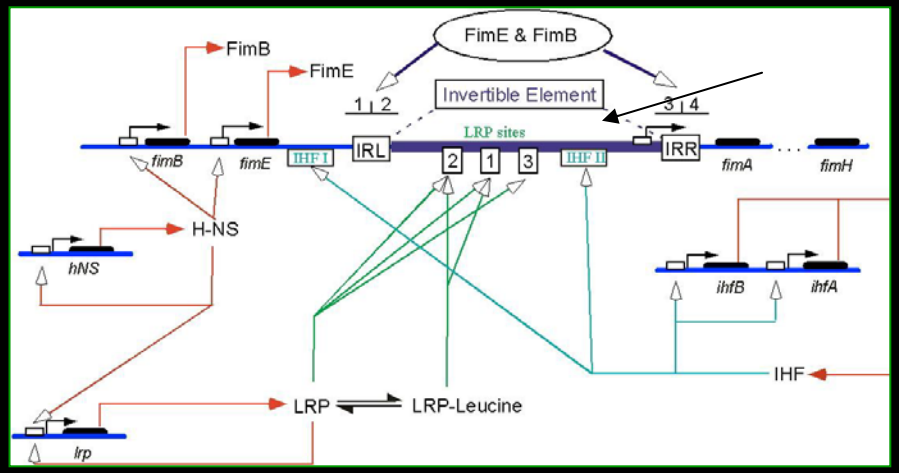
One of the Simplest Genetic Switches



Type-1 Pili Phase Variation in Pathogenic *E. coli*



The Modeling Process: Bottom Up!



This pathway is diagramed using a notation. Stochastic models are chosen for the DNA, and deterministic models for the proteins.

Equilibrium statistical thermodynamics for f, g

$$
\begin{aligned}
dP_{on}/dt &= f*(1-P_{on}) - g*P_{on} \\
&= \frac{\sum_{s\in OFF}\alpha_s e^{-\Delta G_s/RT}[IHF][FimE]^{j(s)}[FimB]^{k(s)}[Lrp*]^{m(s)}[Lrp]^{l(s)}}{1+\sum_{s\in OFF} e^{-\Delta G_s/RT}[IHF][FimE]^{j(s)}[FimB]^{k(s)}[Lrp*]^{m(s)}[Lrp]^{l(s)}}(1-P_{on}) \\
&- \frac{\sum_{s\in ON}\alpha_s e^{-\Delta G_s/RT}[IHF][FimE]^{j(s)}[FimB]^{k(s)}[Lrp*]^{m(s)}[Lrp]^{l(s)}}{1+\sum_{s\in ON} e^{-\Delta G_s/RT}[IHF][FimE]^{j(s)}[FimB]^{k(s)}[Lrp*]^{m(s)}[Lrp]^{l(s)}}P_{on} \qquad (1)
\end{aligned}
$$

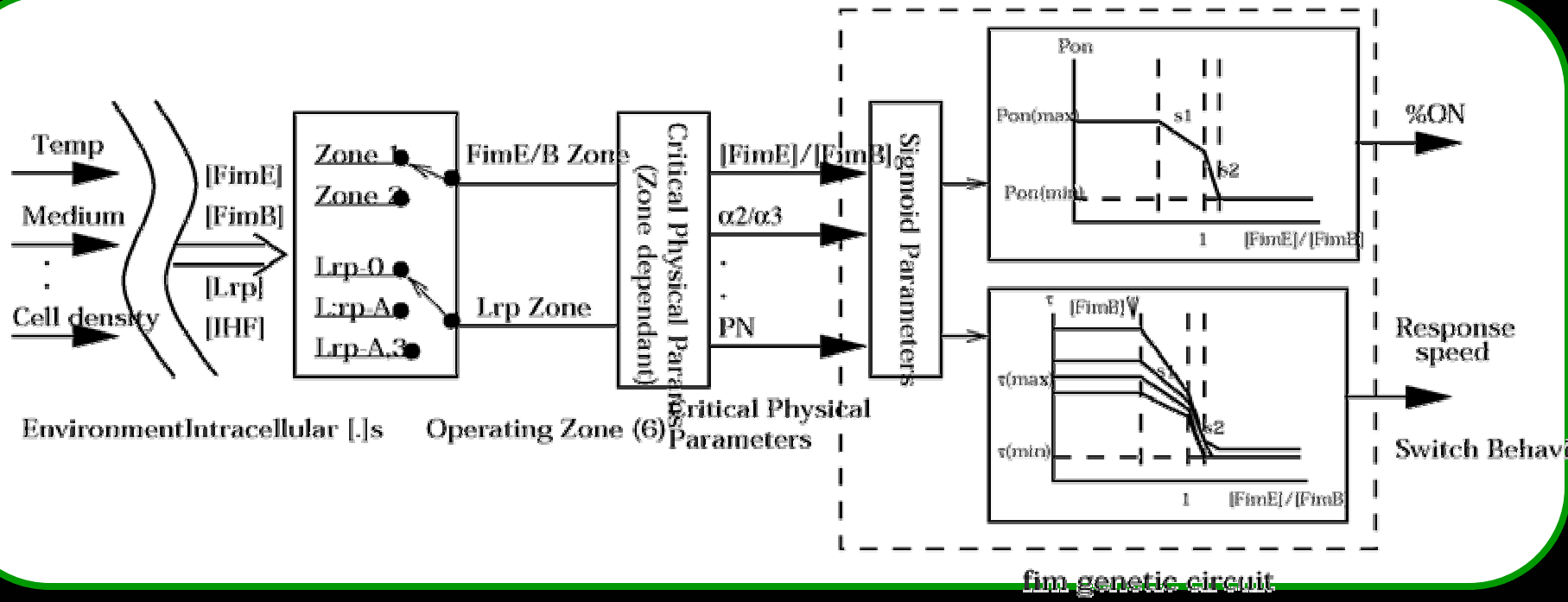
$$
d[FimB]/dt = k_1\frac{e^{-\Delta G_b/RT}[RNAP]}{1+e^{-\Delta G_b/RT}[RNAP]+e^{-\Delta G_{hb}/RT}[H-NS]} - \lambda_1[FimB] \qquad (2)
$$

Protein degradation

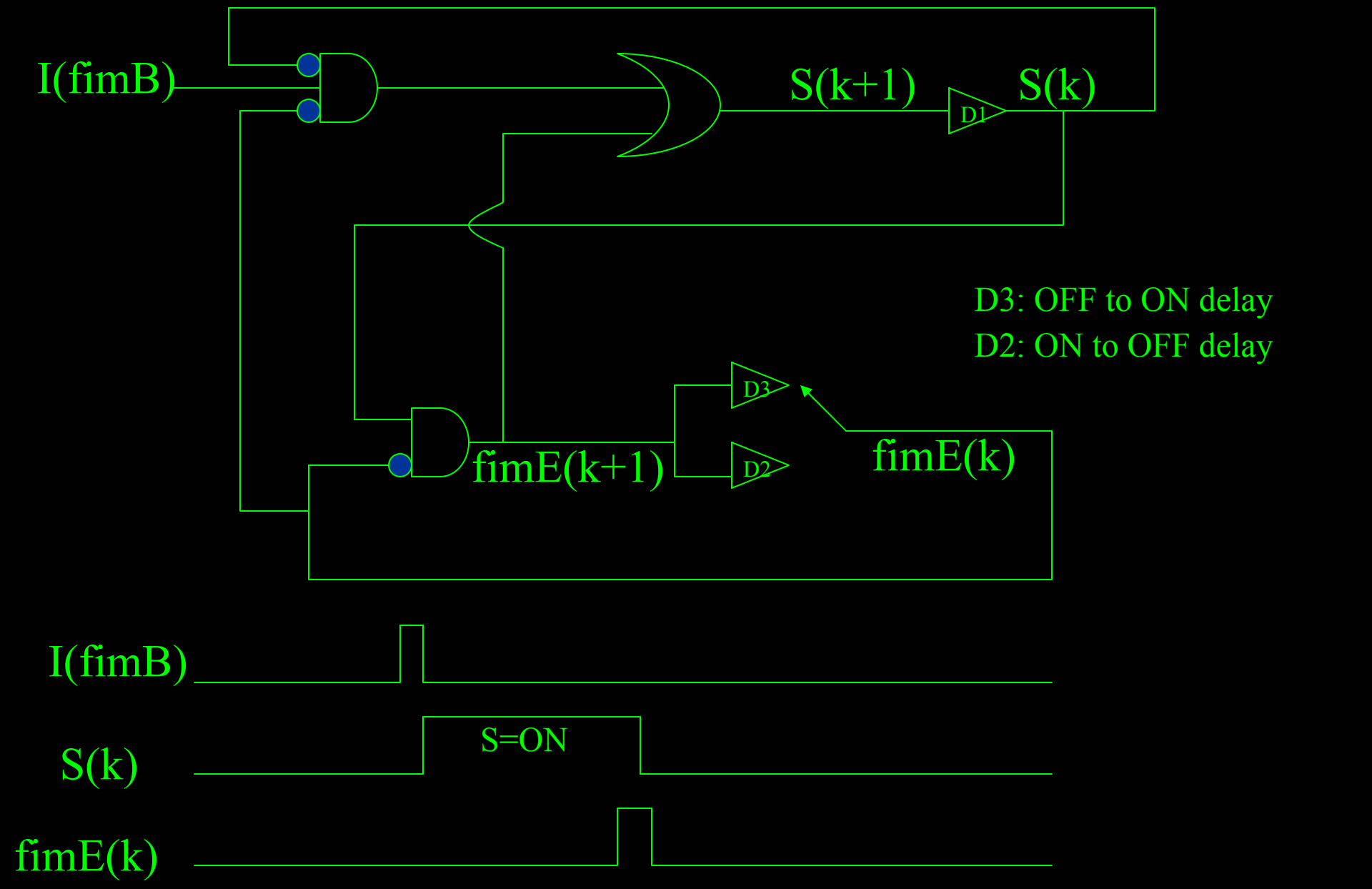
$$
\begin{aligned}
d[FimE]/dt &= (k_2\frac{e^{-\Delta G_e/RT}[RNAP]}{1+e^{-\Delta G_e/RT}[RNAP]+e^{-\Delta G_{he}/RT}[H-NS]})P_{on} \qquad (3) \\
&+ (k_3\frac{e^{-\Delta G_e/RT}[RNAP]}{1+e^{-\Delta G_e/RT}[RNAP]+e^{-\Delta G_{he}/RT}[H-NS]})(1-P_{on}) - \lambda[FimE]
\end{aligned}
$$

This is a hybrid model at the molecular level of abstraction. This now becomes a submodel of a larger infection model.

Expert Logical Abstraction



Any expert examining these results can rapidly deduce a logical control diagram. This is not a formal abstraction but could be used as the basis for a simpler (lower-resolution) model of the switch to be used in the larger simulation.


I(fimB)
S(k+1)
S(k)
D1
D3: OFF to ON delay
D2: ON to OFF delay
D3
D2
fimE(k+1)
fimE(k)
I(fimB)
S=ON
S(k)
fimE(k)

Since biological models must represent models at these different levels of abstraction

How we represent molecules is VERY important.

Either we store models as unrelated directly to primary data in the database as a lump of variables and equations

- Hard to modify
- Hard to relate to data
- Hard to deal with a family of models
- Hard to deal with a linked set of abstractions (stay tuned)

Or we ensure that the objects that models describe are represented for modeling.

If a phosophylatable protein is one molecule with internal state, how does a model specification refer to it?

If we represent all states the DB gets bloated?

If states group how do we group them?

Adam's ignorant scheme

1. If data exists on a particular state of a molecule it is given its own record
2. This record is referenced back to the "parent" molecules
    1. Defined as the molecules from which this molecule may be created.
3. If data on a particular "possible" state does not exist do the same thing.
4. Define equivalence class specifications
    1. This is a model object!

Representation and interoperability: Models can be passed around

CellML- http://www.cellml.org

SBML- http://www.cds.caltech.edu/erato/

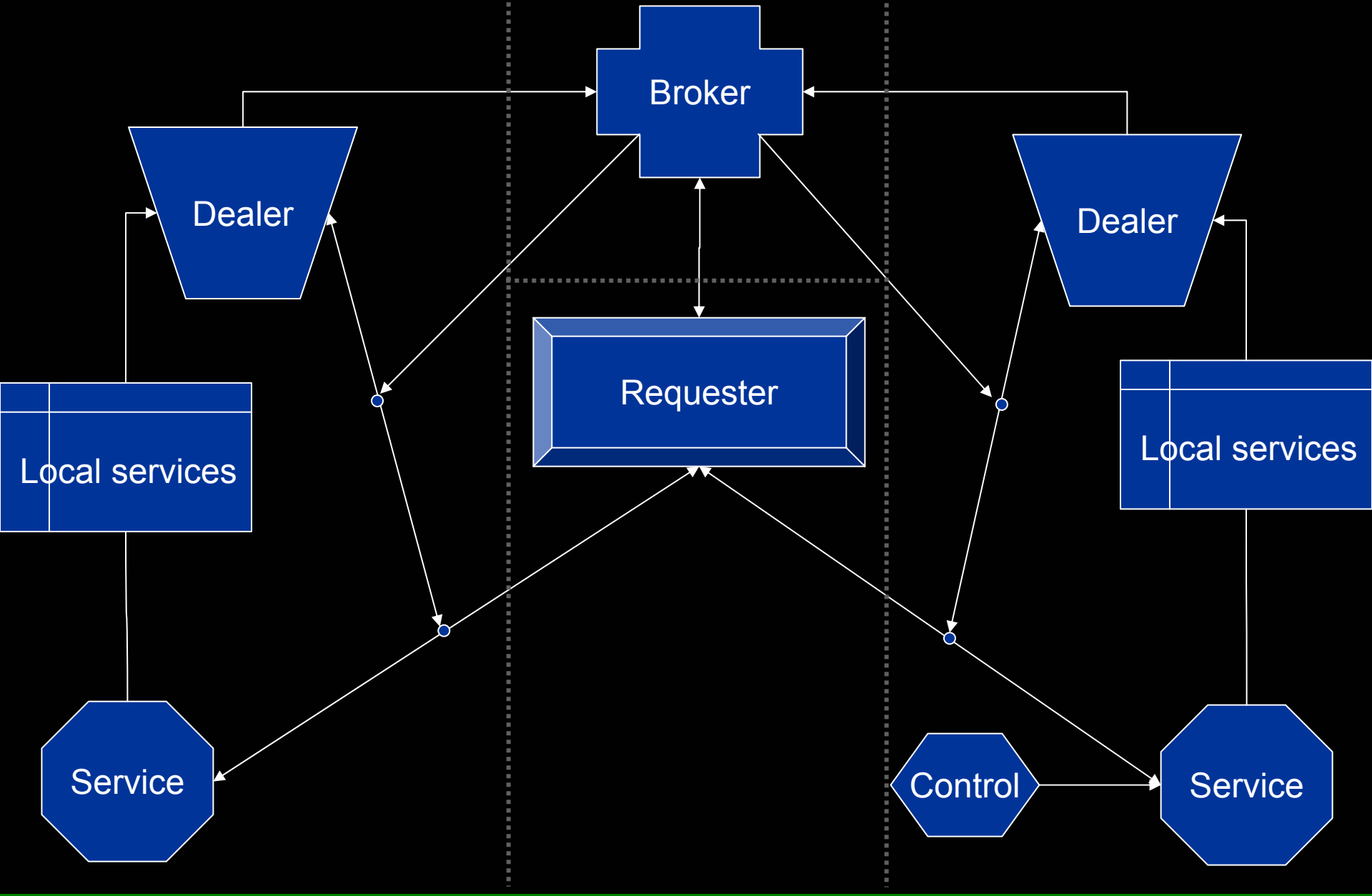
JOIN THE DISCUSSION NOW!

```
<model>
    ...
    <listOfUnitDefinitions>
        <unitDefinition name="volume">
            <listOfUnits>
                <unit kind="liters" scale="-3"/>
            </listOfUnits>
        </unitDefinition>
    </listOfUnitDefinitions>
    ...
</model>
```

These specifications are designed for sending models to simulators.
But what if there is data comparison, etc.

However, this is “middleware” and can solve a plethora of problems

Distributed, loosely coupled, architecture



Conclusions

- Data representation is the lowest level of representation of biological knowledge
- Models are particular “statements” of this knowledge.
- Databases and models must be linked for comparison
- How schema should be designed to facilitate this is research
  - Much can be fixed post-facto with “middleware”

Data analysis, Modules and Models

Prospectus and Problems

Pessimistic optimism

Panglossian Pessimism

Data analysis

An Example of Effect Detection

Questions

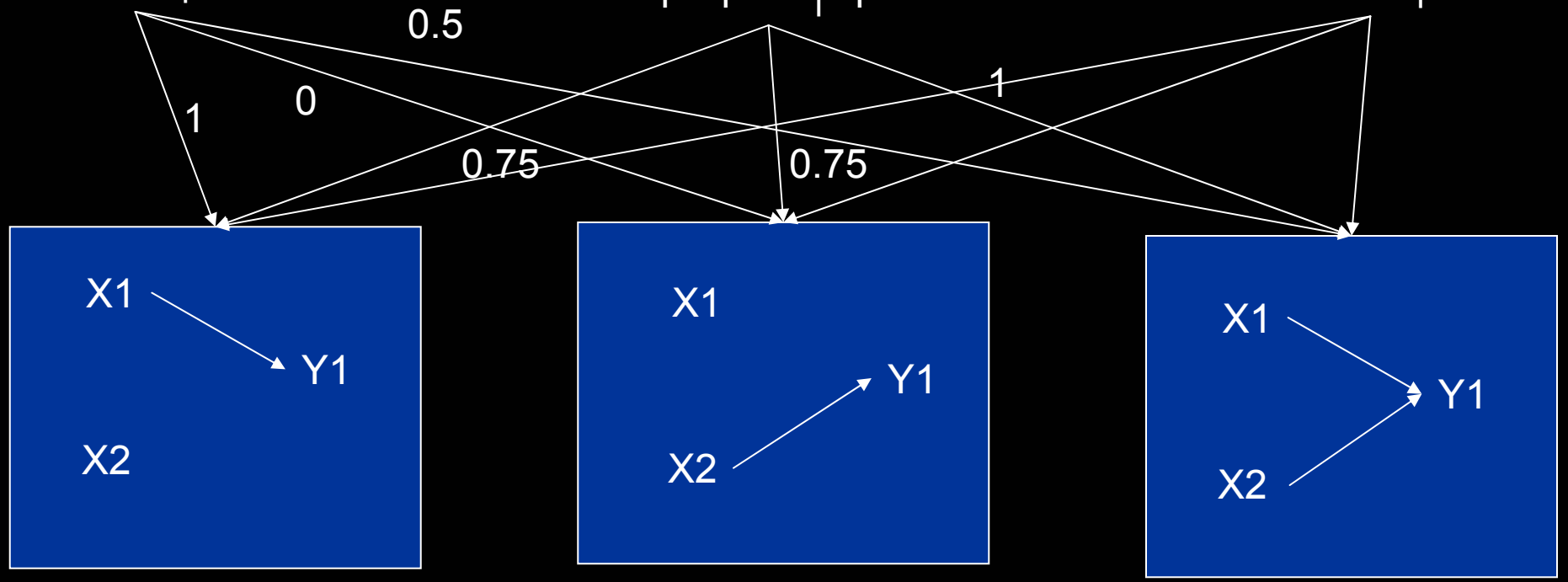
- What's my question?
  - Does this perturbation have an effect?
  - Is loss of this interaction responsible for X?
  - Can the known network reproduce dynamics?
- What experimental design best answers my question?
- Do I need complicated statistics?

Graphical Models: An incredibly stupid example!

X1	X2	Y1
0	0	1
0	1	1
1	0	0
1	1	0

X1	X2	Y1
0	0	0
0	1	1
1	0	1
1	1	1

X1	X2	Y1
0	0	0
0	1	0
1	0	0
1	1	1



Graphical Models: An incredibly stupid example!

X1	X2	Y1
0	0	1
0	1	1
1	0	0
1	1	0

X1	X2	Y1
0	0	0
0	1	1
1	0	1
1	1	1

X1	X2	Y1
0	0	0
0	1	0
1	0	0
1	1	1

For simple discrete *combinational* data, we can enumerate all possible states and deduce directly.

For sequential data, large combinational data or continuous data things become more complicated.

We need roughly measures of the sort:

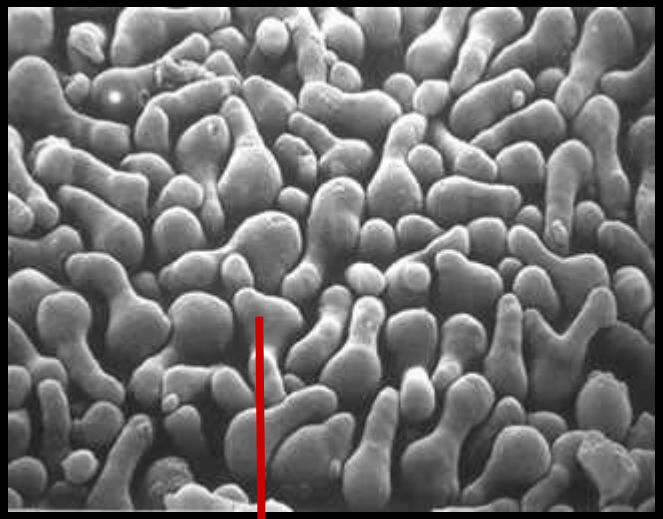
$$P(Y=y \mid X1,X2,\ldots)$$

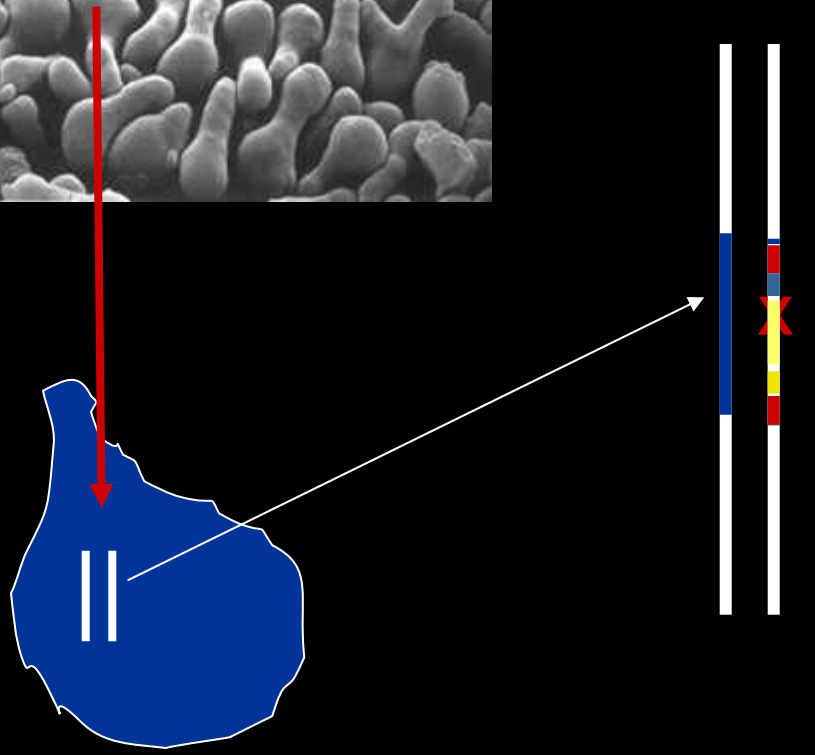
Simpler Question

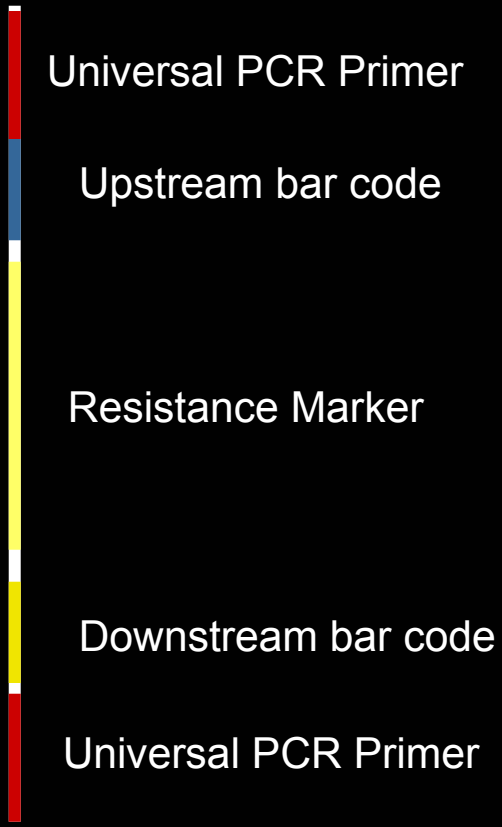
Effect Detection
Guri Giaever and Ron Davis

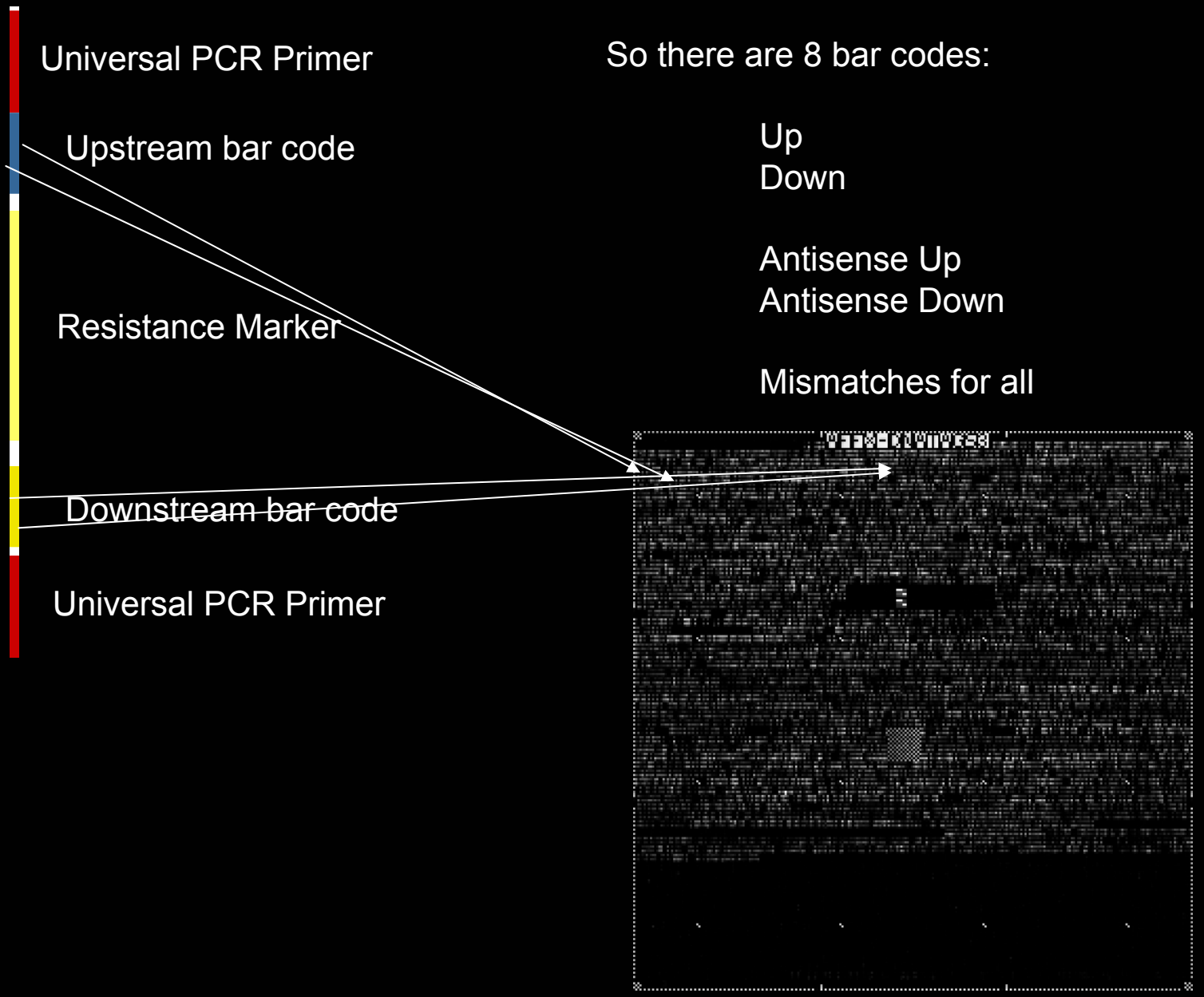
(confidential and VERY early)

Yeast Haploinsufficiency Trials









So there are 8 bar codes:

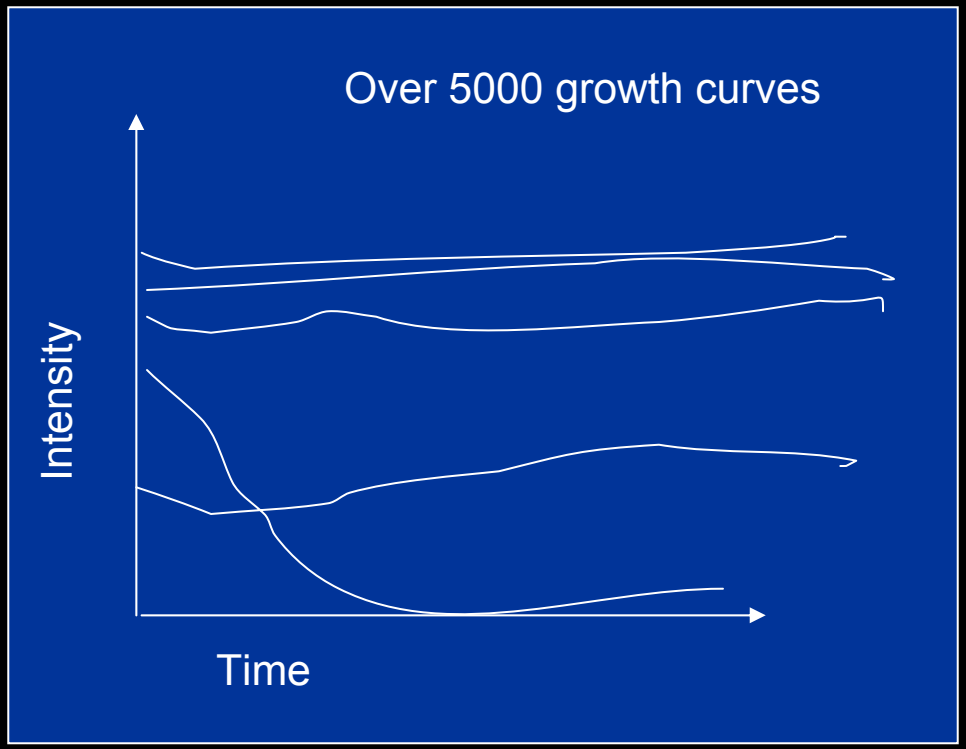
Up
Down

Antisense Up
Antisense Down

Mismatches for all

Yeast Haploinsufficiency Trials

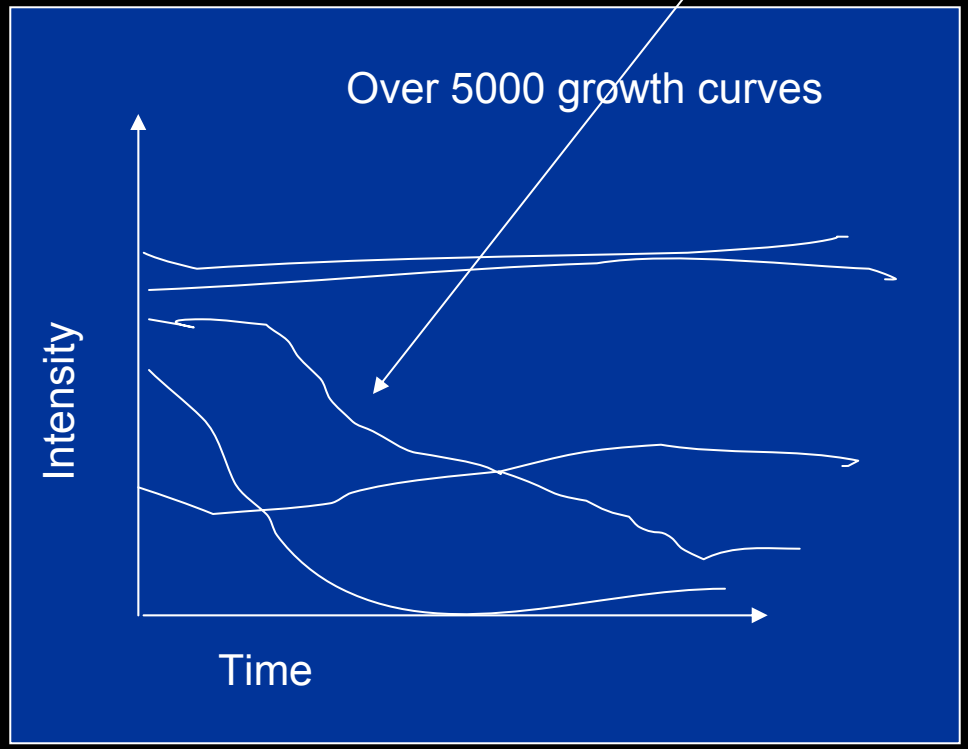
Sample every 0.5 population doubling.
Dilute sample to standard OD.
Examine on chip.



Yeast Haploinsufficiency Trials

Drop drug in!

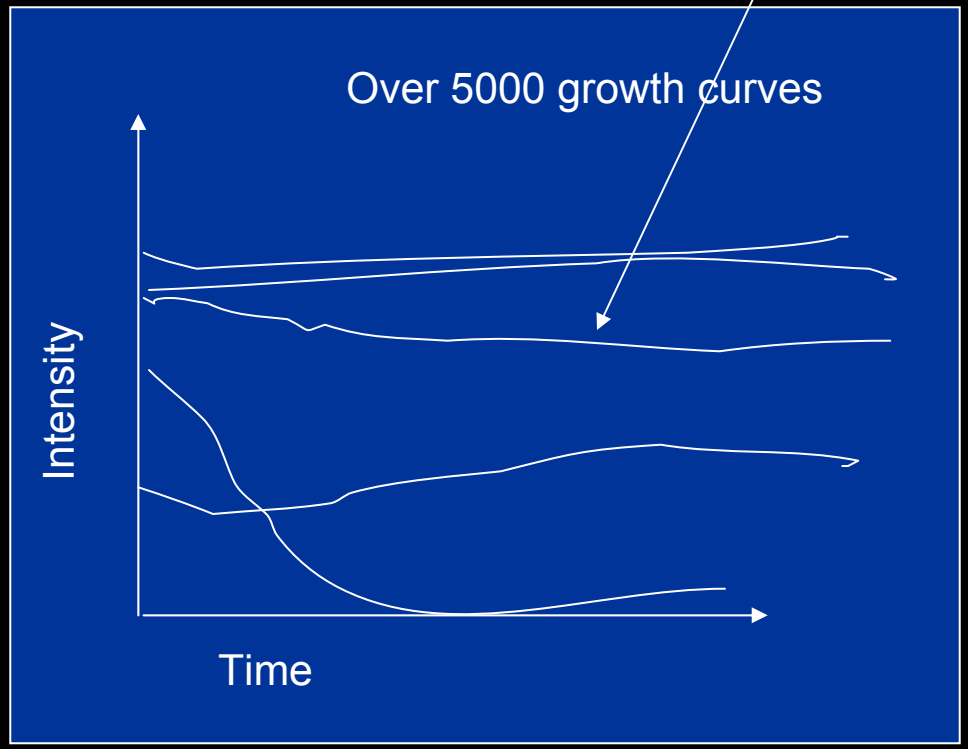
This guy now drops out!



Yeast Haploinsufficiency Trials

Drop drug in!

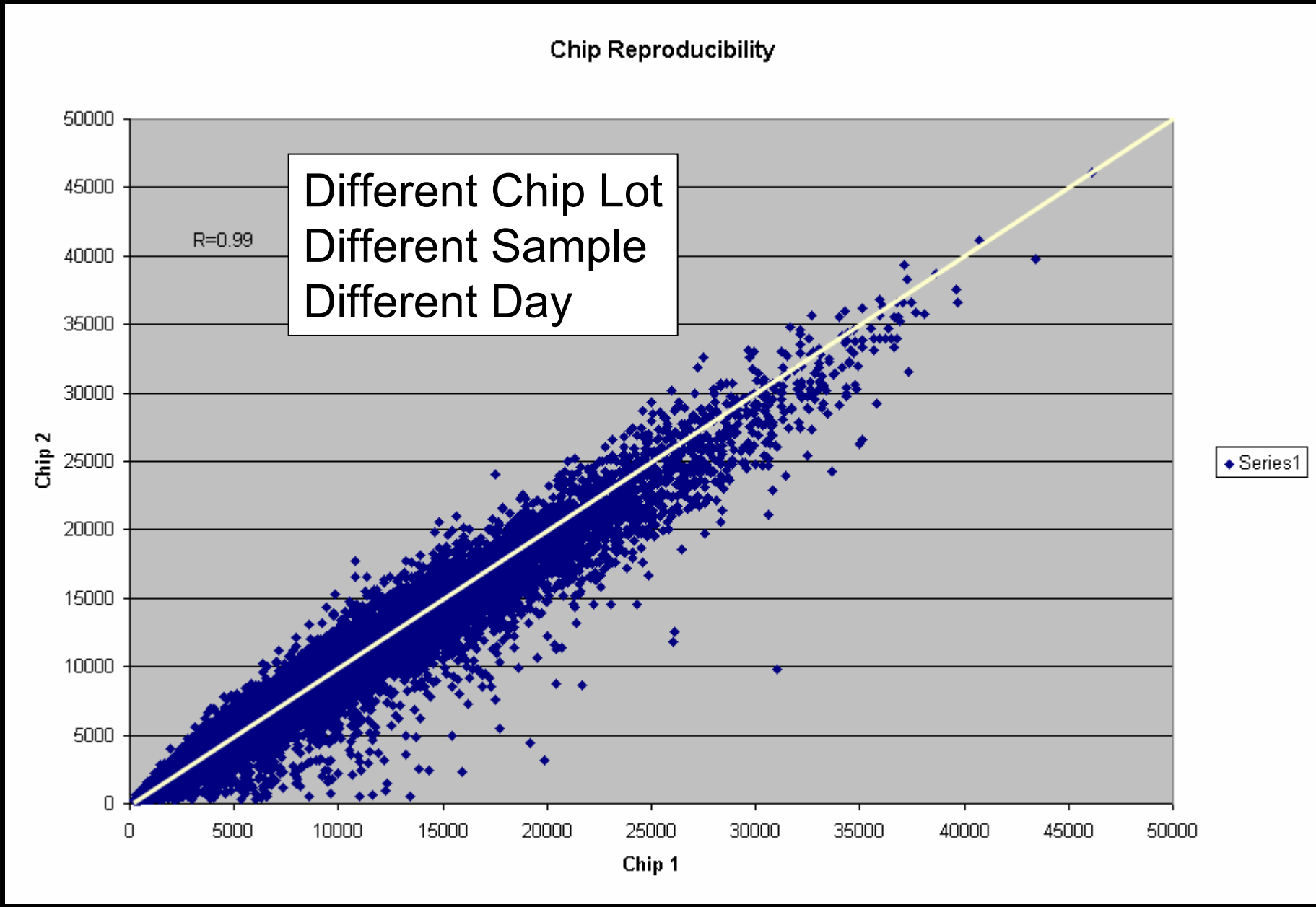
Is this guy dropping out?

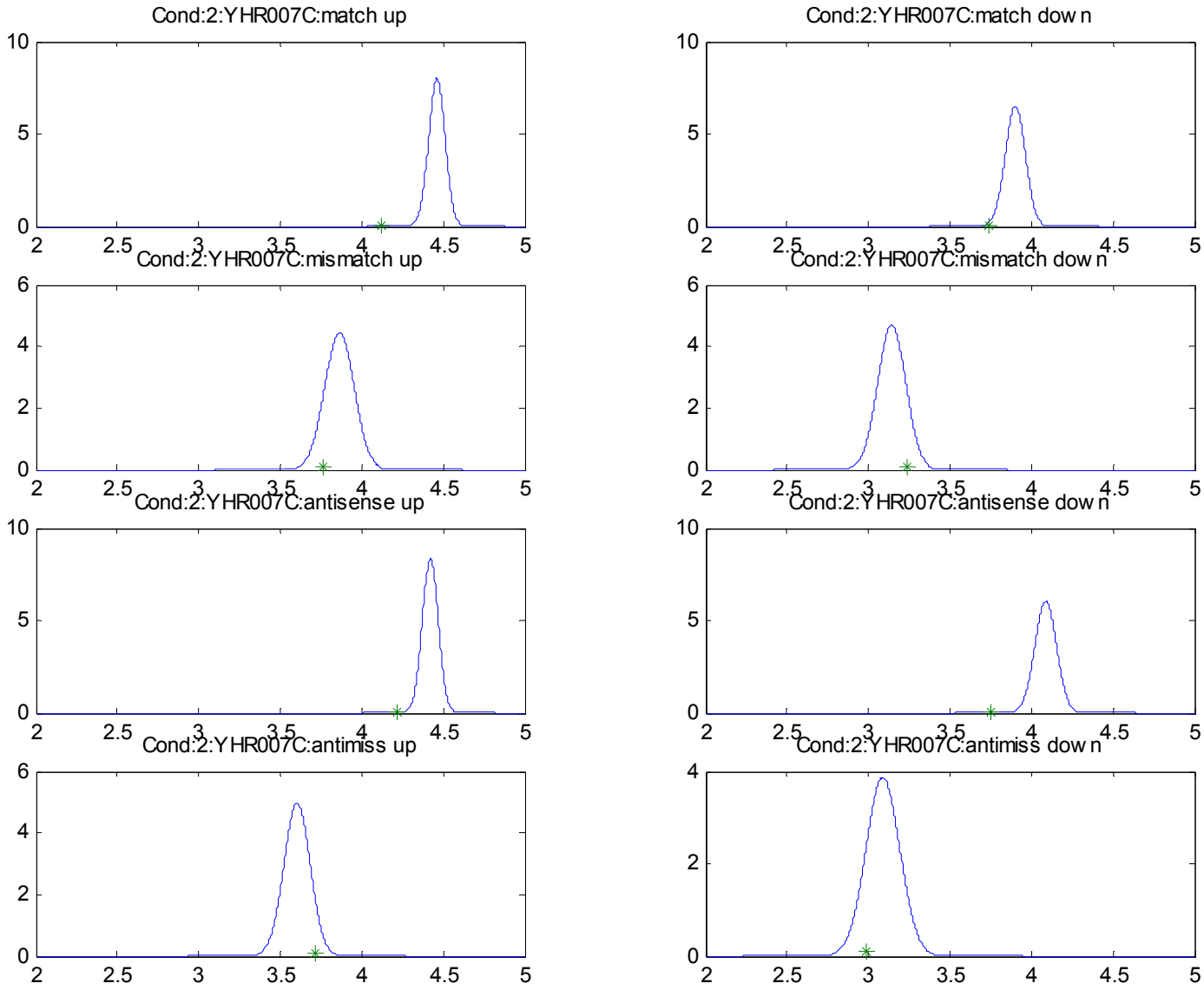


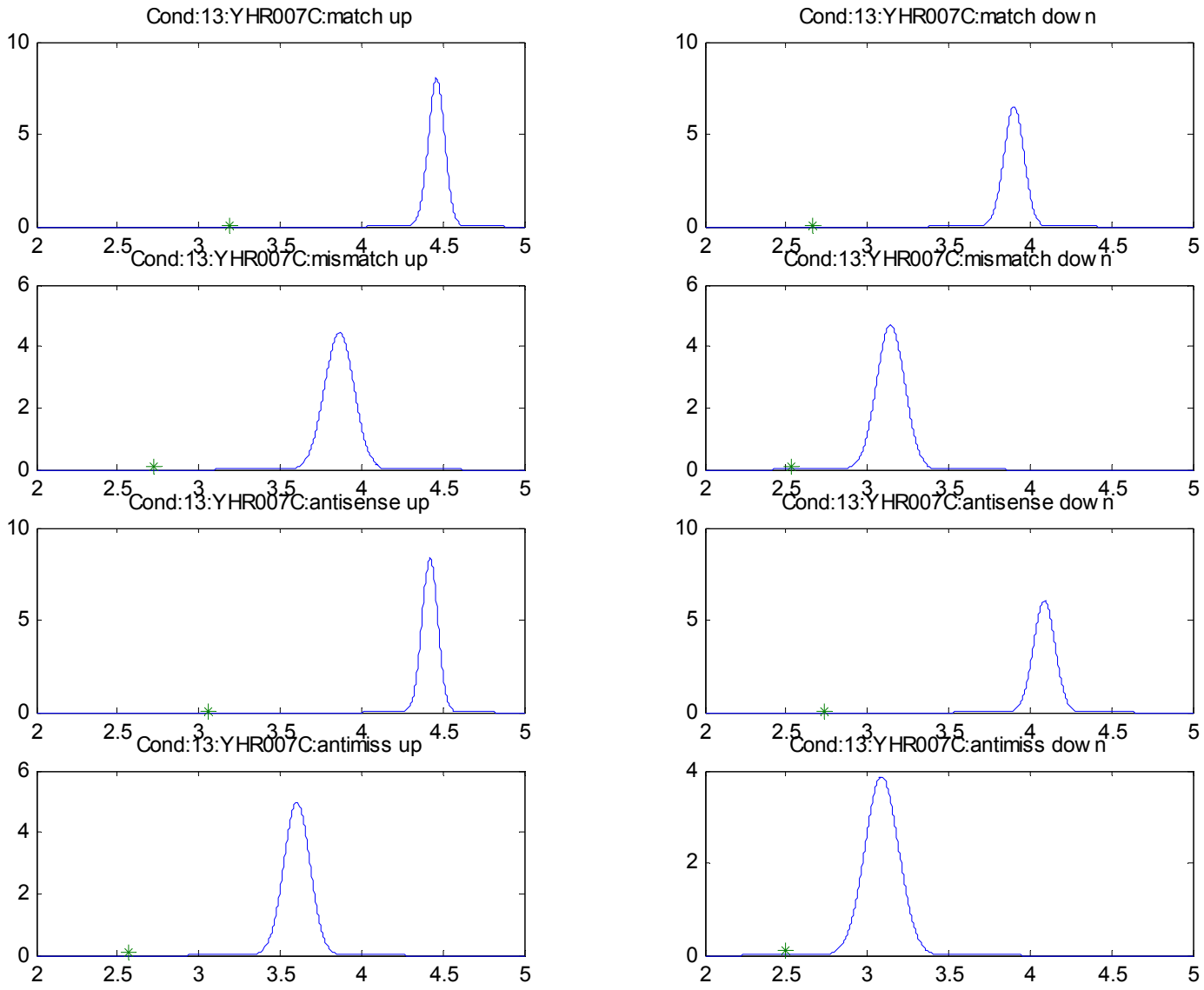
Approach

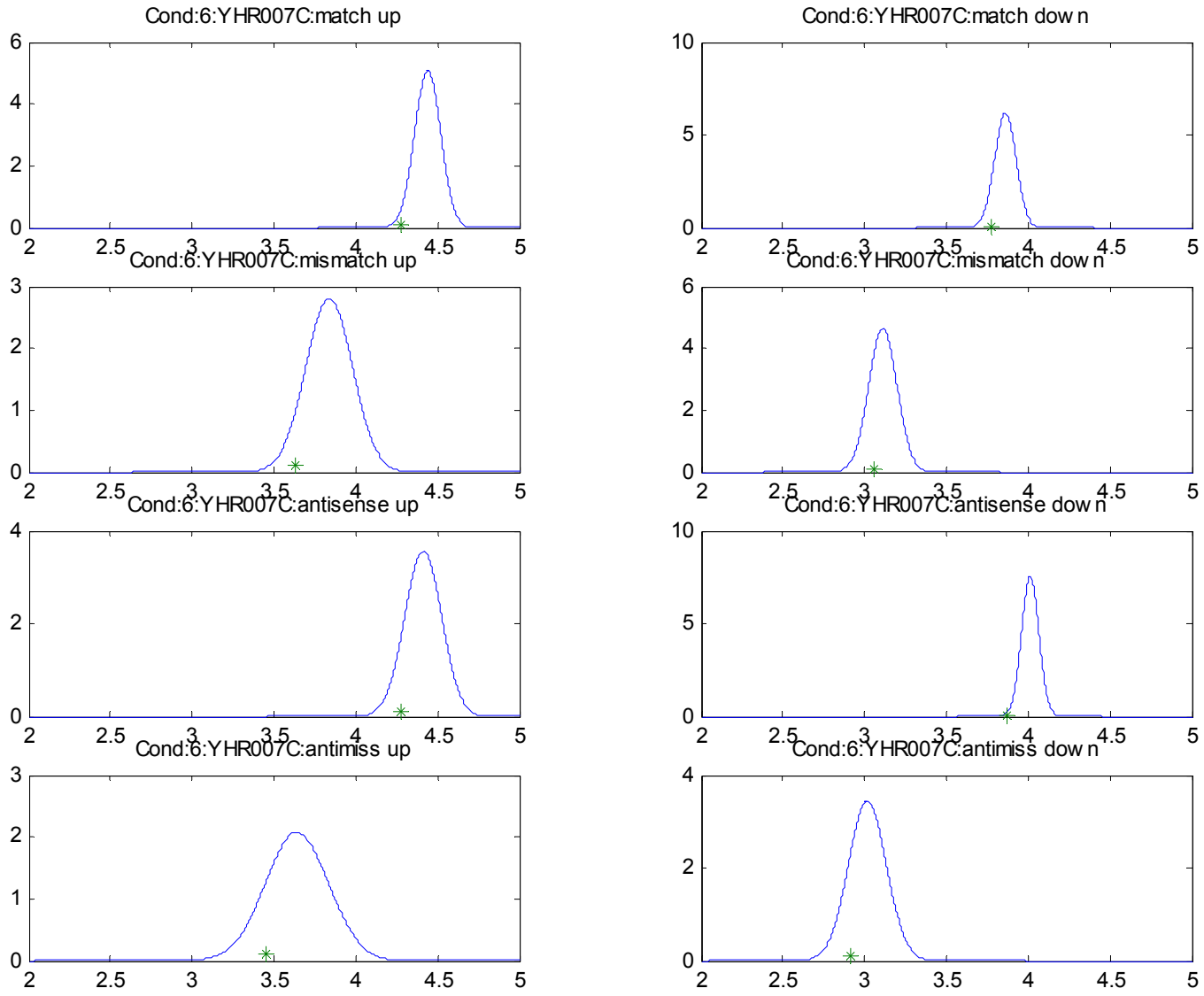
1. Look at time course, fit to growth curve, score by decay
2. Look at time zero, look at time=T, score by ratio

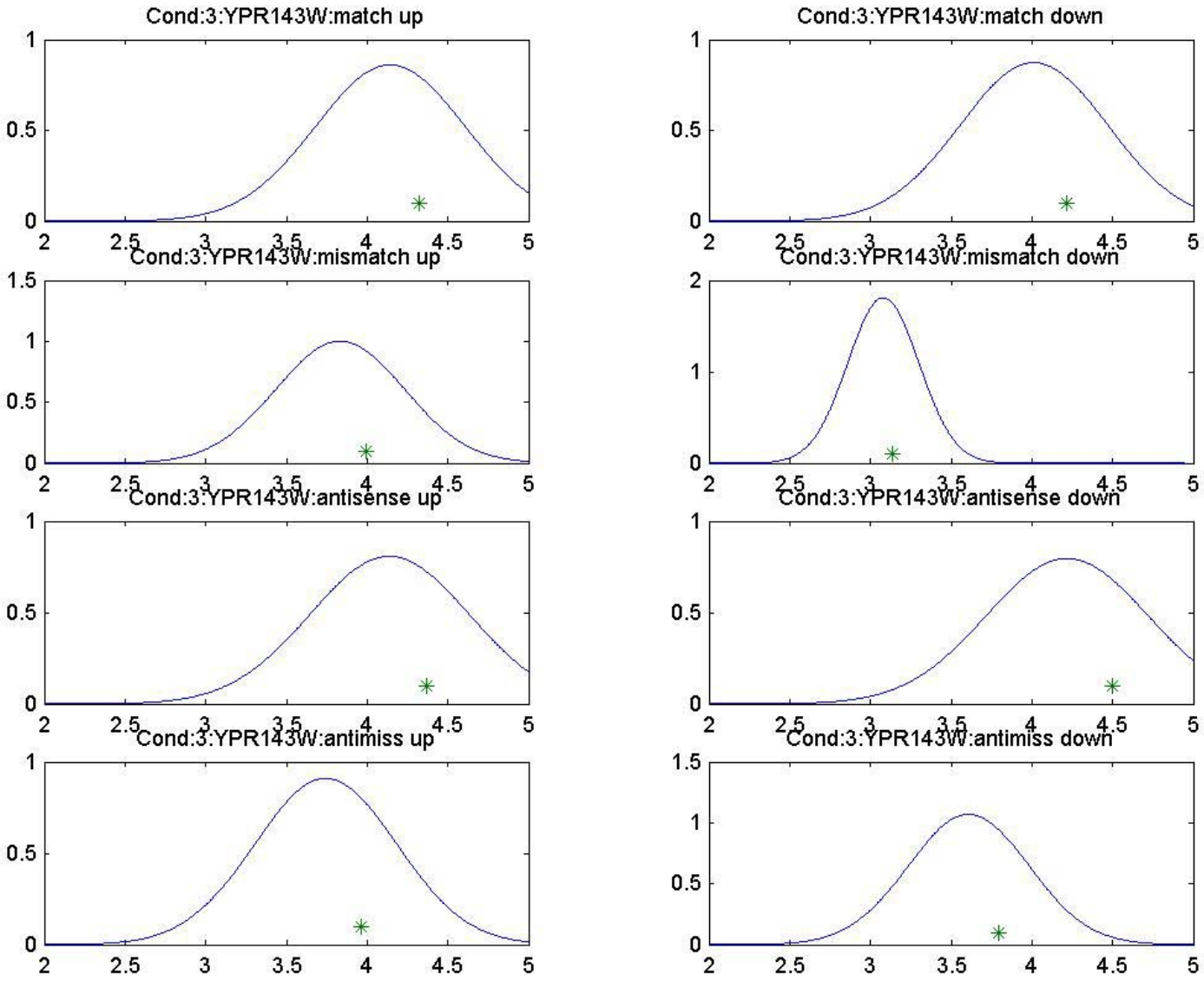
3. Develop likelihood measure. (A stupid but effective one)
    1. Decide on a standard condition
    2. Do N replicates at each generation time
    3. Estimate the distribution of intensity for each tag at a given time.
    4. Score new experiments as the probability of being drawn from that distribution.

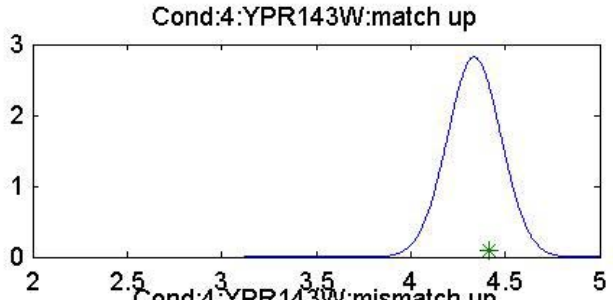

Chip Reproducibility
R=0.99
Different Chip Lot
Different Sample
Different Day
Chip 2
Chip 1
Series1
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000

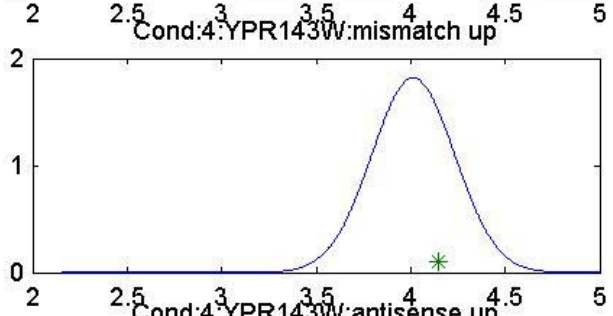

Cond:2:YHR007C:match up
Cond:2:YHR007C:match down
Cond:2:YHR007C:mismatch up
Cond:2:YHR007C:mismatch down
Cond:2:YHR007C:antisense up
Cond:2:YHR007C:antisense down
Cond:2:YHR007C:antimiss up
Cond:2:YHR007C:antimiss down

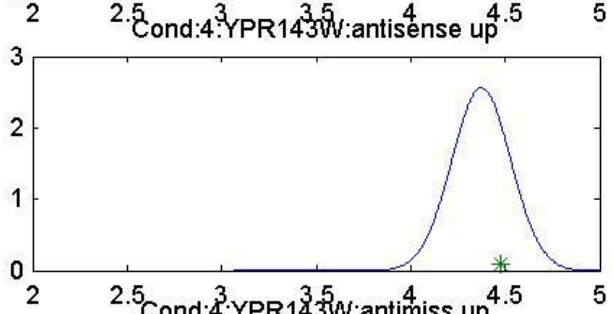

Cond:13:YHR007C:match up
Cond:13:YHR007C:match down
Cond:13:YHR007C:mismatch up
Cond:13:YHR007C:mismatch down
Cond:13:YHR007C:antisense up
Cond:13:YHR007C:antisense down
Cond:13:YHR007C:antimiss up
Cond:13:YHR007C:antimiss down

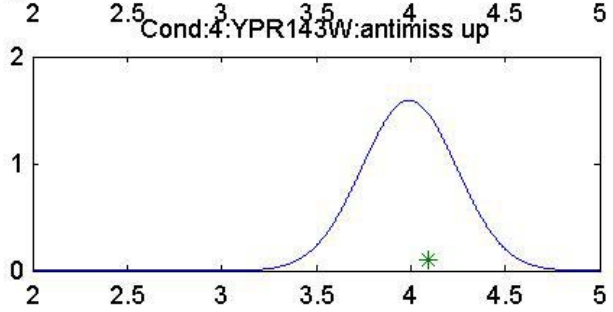

Cond:6:YHR007C:match up
Cond:6:YHR007C:match down
Cond:6:YHR007C:mismatch up
Cond:6:YHR007C:mismatch down
Cond:6:YHR007C:antisense up
Cond:6:YHR007C:antisense down
Cond:6:YHR007C:antimiss up
Cond:6:YHR007C:antimiss down

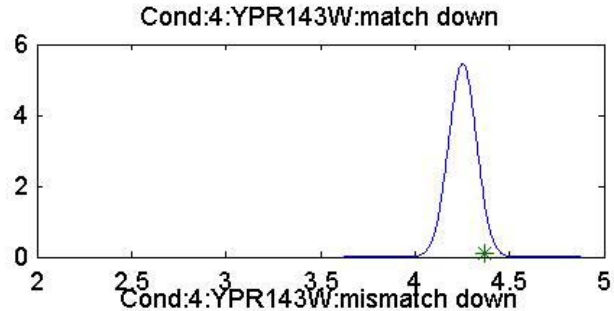

Cond:3:YPR143W:match up
Cond:3:YPR143W:match down
Cond:3:YPR143W:mismatch up
Cond:3:YPR143W:mismatch down
Cond:3:YPR143W:antisense up
Cond:3:YPR143W:antisense down
Cond:3:YPR143W:antimiss up
Cond:3:YPR143W:antimiss down

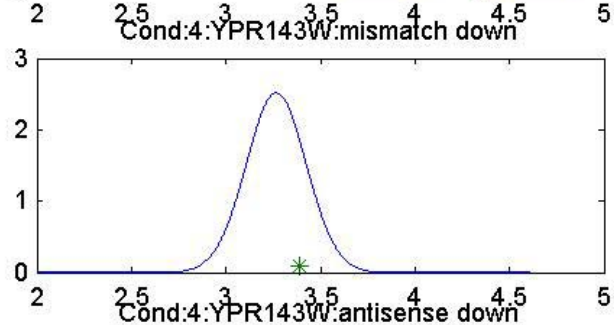

Cond:4:YPR143W:match up
3
2
1
0
2
2.5
3
3.5
4
4.5
5

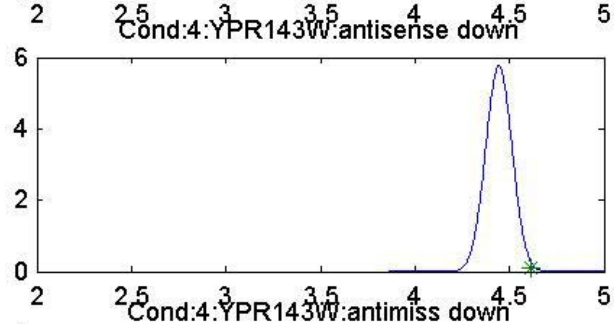

Cond:4:YPR143W:mismatch up
2
1
0
2
2.5
3
3.5
4
4.5
5

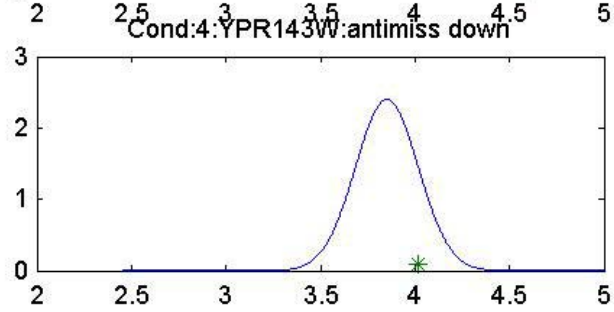

Cond:4:YPR143W:antisense up
3
2
1
0
2
2.5
3
3.5
4
4.5
5

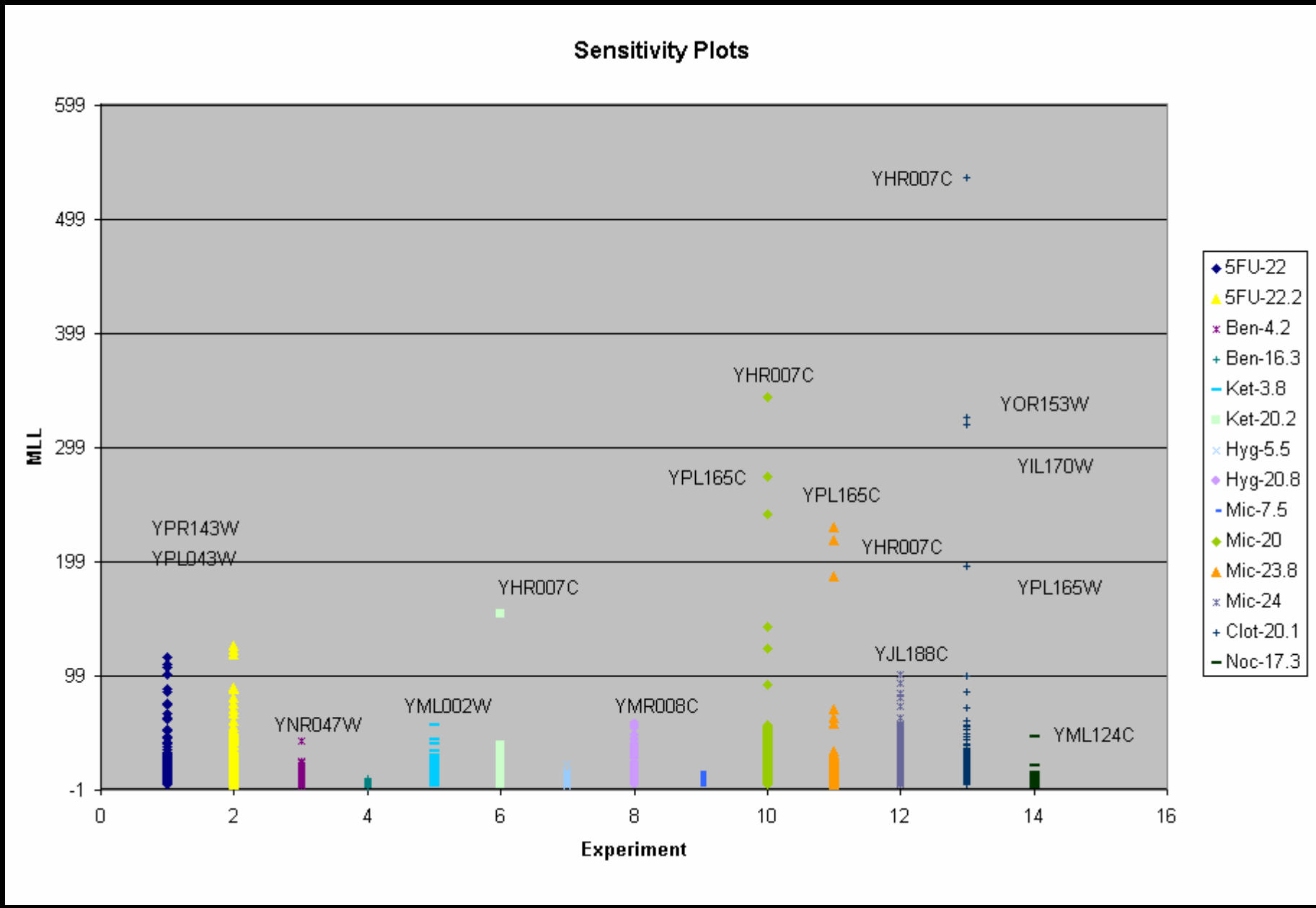

Cond:4:YPR143W:antimiss up
2
1
0
2
2.5
3
3.5
4
4.5
5


Cond:4:YPR143W:match down
6
4
2
0
2
2.5
3
3.5
4
4.5
5


Cond:4:YPR143W:mismatch down
3
2
1
0
2
2.5
3
3.5
4
4.5
5


Cond:4:YPR143W:antisense down
6
4
2
0
2
2.5
3
3.5
4
4.5
5


Cond:4:YPR143W:antimiss down
3
2
1
0
2
2.5
3
3.5
4
4.5
5


Sensitivity Plots
599
499
399
299
199
99
-1
MLL
0
2
4
6
8
10
12
14
16
Experiment
YHR007C
YHR007C
YOR153W
YIL170W
YPL165C
YPL165C
YPR143W
YPL043W
YHR007C
YHR007C
YPL165W
YJL188C
YML002W
YMR008C
YNR047W
YML124C
5FU-22
5FU-22.2
Ben-4.2
Ben-16.3
Ket-3.8
Ket-20.2
Hyg-5.5
Hyg-20.8
Mic-7.5
Mic-20
Mic-23.8
Mic-24
Clot-20.1
Noc-17.3

Experiment 6: Ket 20.2 hours

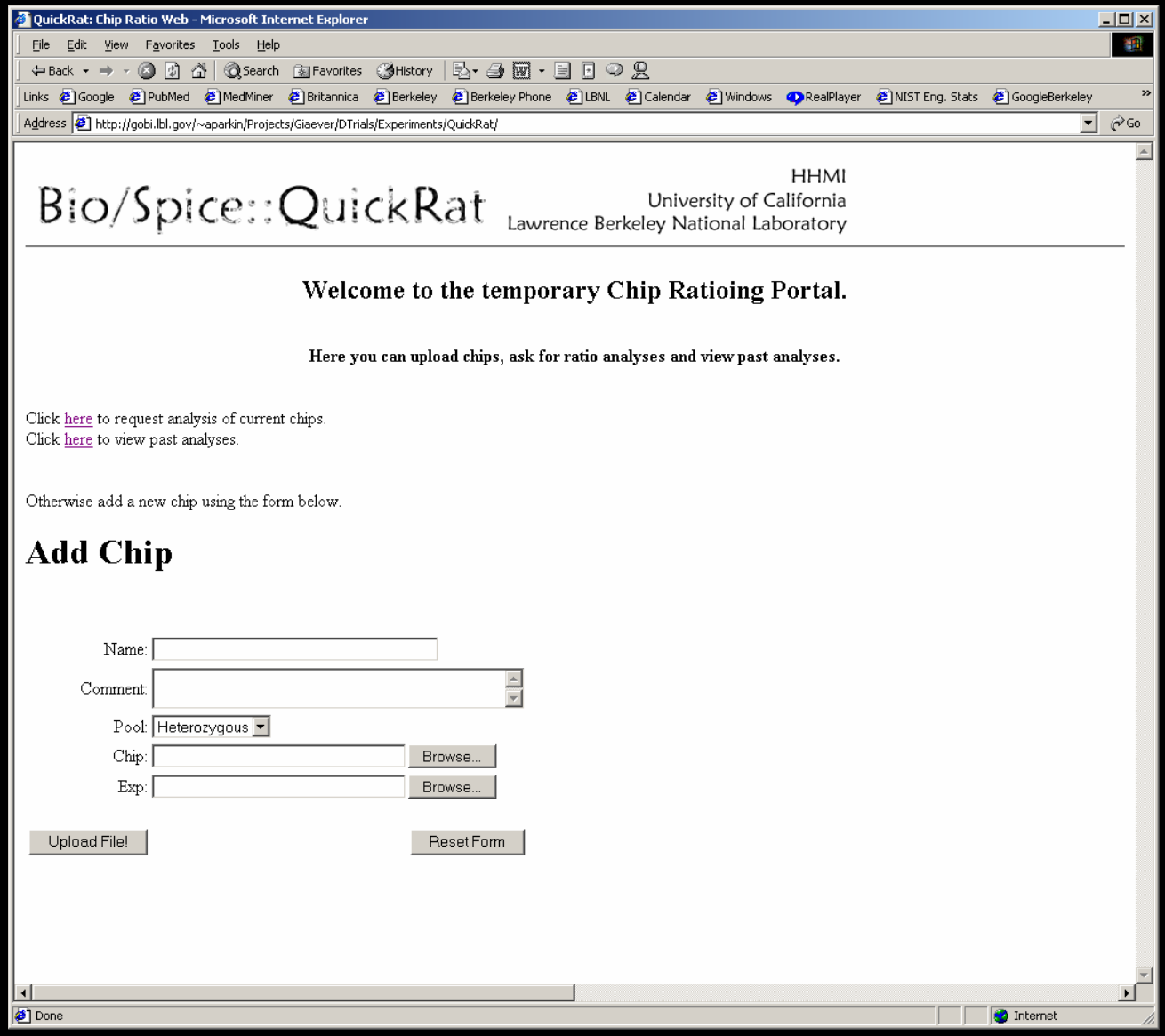
YHR007C	ERG11	cytochrome P450 lanosterol 14a-demethylase, ergosterol biosynthesis, lanosterol 14-alpha-demethylase, endoplasmic reticulum
YML085C	Tub1	alpha-tubulin, mitotic chromosome segregation*, structural protein of cytoskeleton, spindle pole body
YJR097W		

Experiment 10: Mic 20 hours		
YHR007C	ERG11	cytochrome P450 lanosterol 14a-demethylase, ergosterol biosynthesis, lanosterol 14-alpha-demethylase, endoplasmic reticulum
YPL165C	UNK	UNK
YIL170W	HXT12	Hexose permease transport
YOR153W	PDR5,LEM1, YDR1	multidrug resistance transporter
YLR208W	Sec13	cytoplasmic protein involved in release of transport vesicles from the ER, non-selective vesicle assembly, molecular_function unknown, cytoplasm
YML085C	Tub1	alpha-tubulin, mitotic chromosome segregation*, structural protein of cytoskeleton, spindle pole body

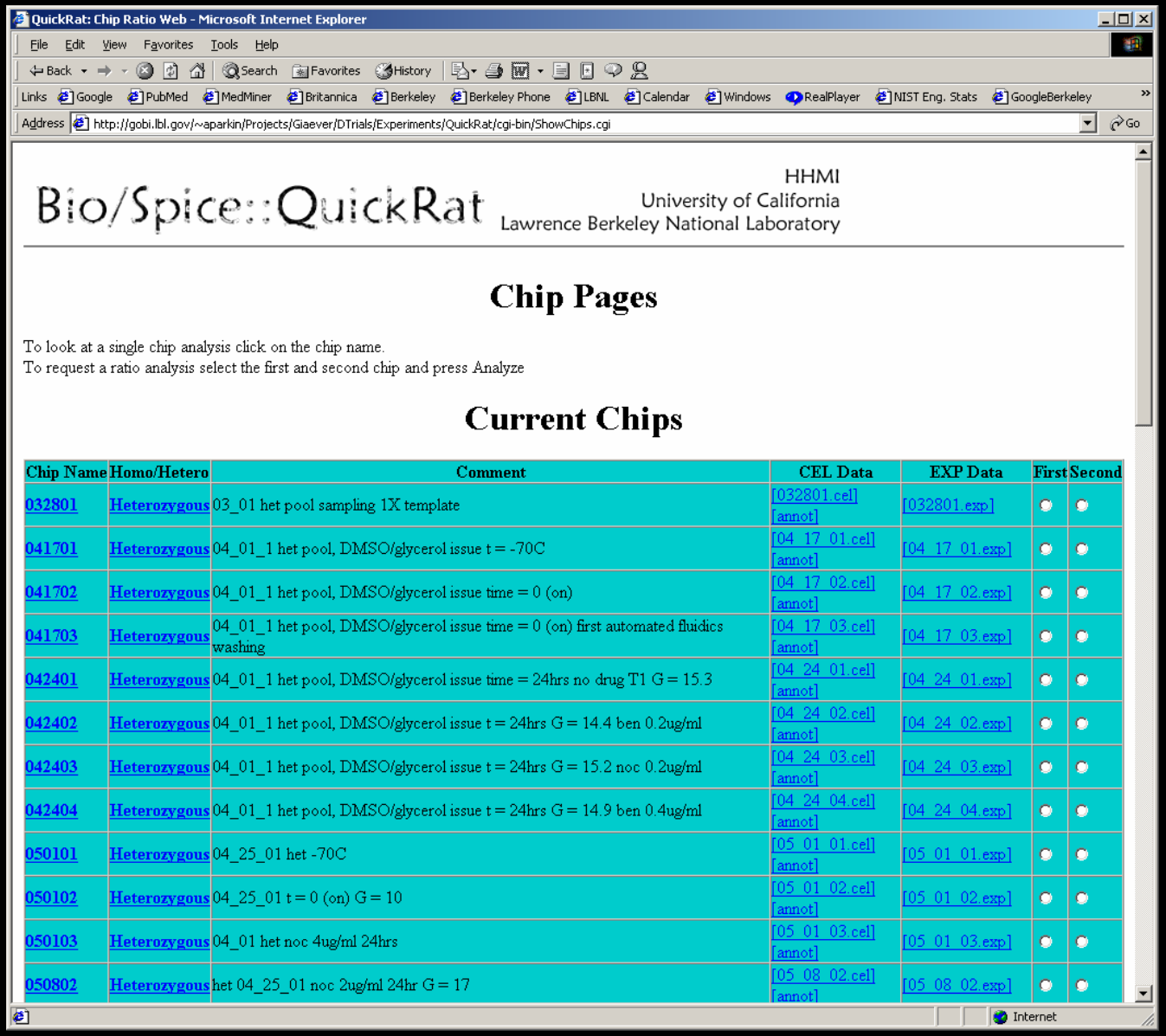
Experiment 13: Clot 20.1 hours		
YHR007C	ERG11	cytochrome P450 lanosterol 14a-demethylase, ergosterol biosynthesis, lanosterol 14-alpha-demethylase, endoplasmic reticulum
YOR153W	PDR5,LEM1, YDR1	multidrug resistance transporter
YPL165C	UNK	UNK
YIL170W	HXT12	Hexose permease transport
YML085C	Tub1	alpha-tubulin, mitotic chromosome segregation*, structural protein of cytoskeleton, spindle pole body

Experiment 14: Noc 17.3 hours		
YML124C	TUB3	alpha-tubulin, mitotic chromosome segregation*, structural protein of cytoskeleton, spindle pole bod
YJL014W	CCT3, BIN2, TCP3	Cytoplasmic chaperonin subunit gamma, protein folding*, chaperone, cytoplasm
YJR097W	RPL27A	Ribosomal protein L27A, protein biosynthesis, structural protein of ribosome, cytosolic large ribosomal (60S)-subunit

Simplified Data Upload

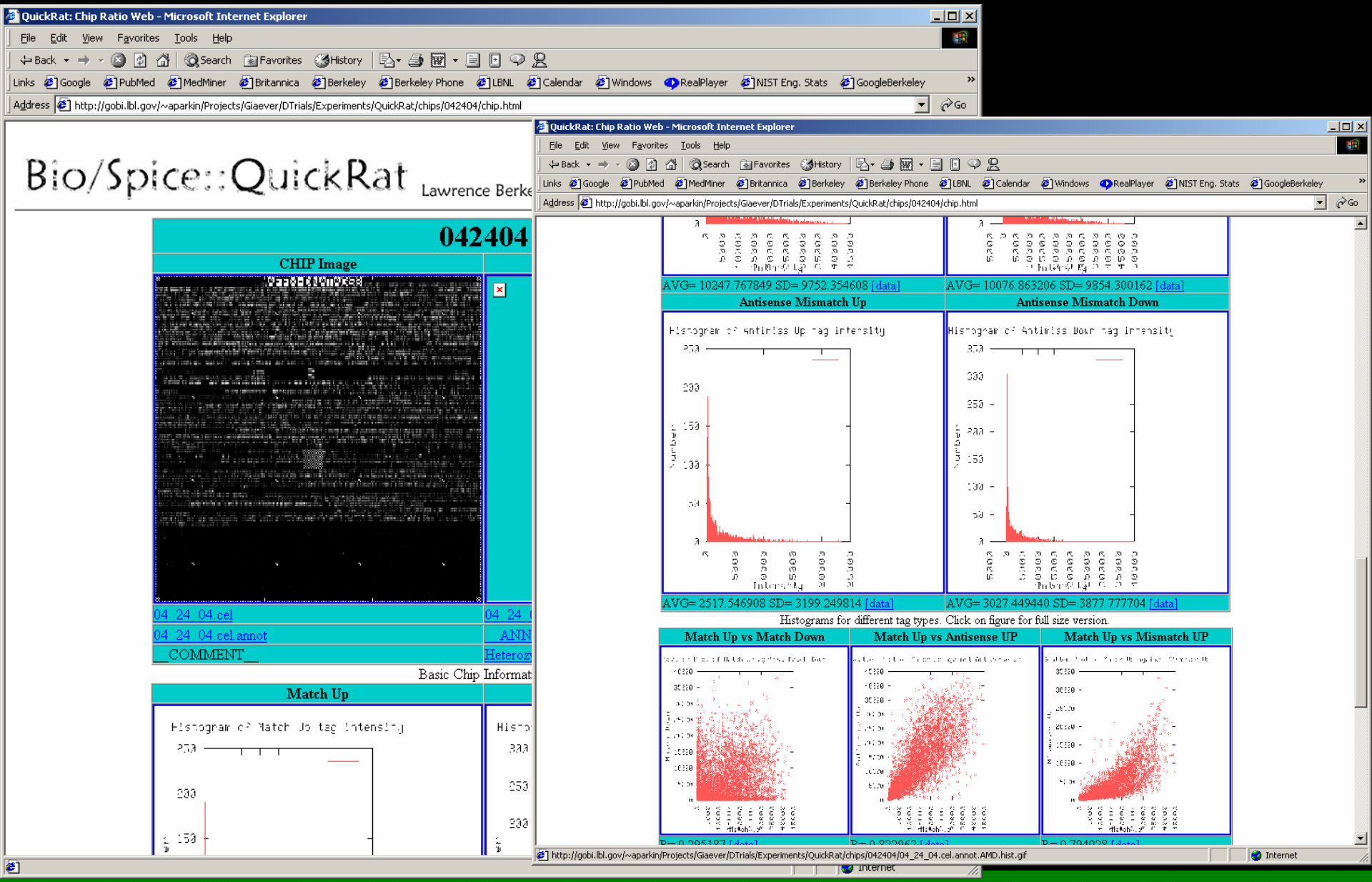


Easy access to chips/requests for analysis

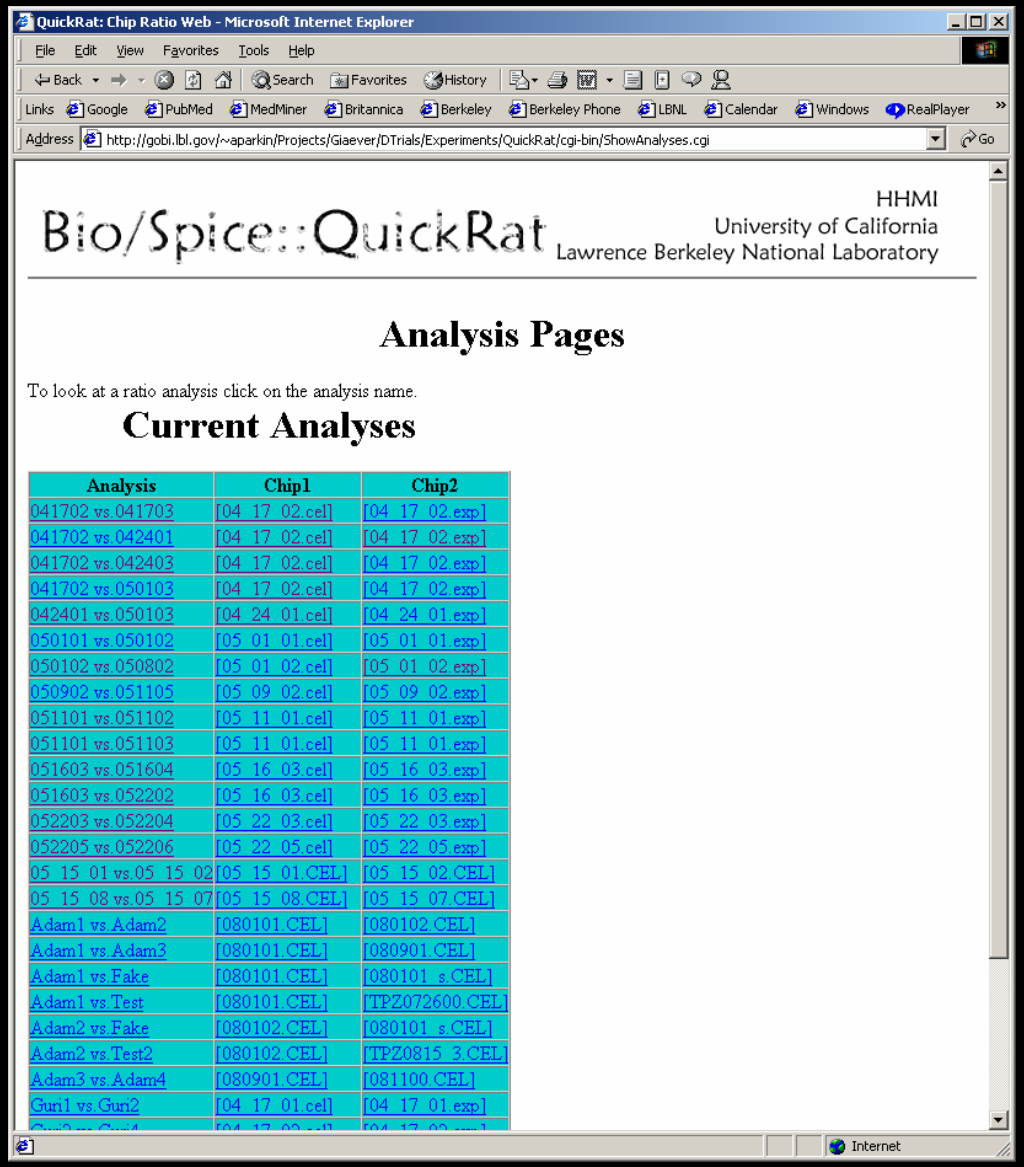


Chip Name	Homo/Hetero	Comment	CEL Data	EXP Data	First	Second
032801	Heterozygous	03_01 het pool sampling 1X template	[032801.cel] [annot]	[032801.exp]		
041701	Heterozygous	04_01_1 het pool, DMSO/glycerol issue t = -70C	[04_17_01.cel] [annot]	[04_17_01.exp]		
041702	Heterozygous	04_01_1 het pool, DMSO/glycerol issue time = 0 (on)	[04_17_02.cel] [annot]	[04_17_02.exp]		
041703	Heterozygous	04_01_1 het pool, DMSO/glycerol issue time = 0 (on) first automated fluidics washing	[04_17_03.cel] [annot]	[04_17_03.exp]		
042401	Heterozygous	04_01_1 het pool, DMSO/glycerol issue time = 24hrs no drug T1 G = 15.3	[04_24_01.cel] [annot]	[04_24_01.exp]		
042402	Heterozygous	04_01_1 het pool, DMSO/glycerol issue t = 24hrs G = 14.4 ben 0.2ug/ml	[04_24_02.cel] [annot]	[04_24_02.exp]		
042403	Heterozygous	04_01_1 het pool, DMSO/glycerol issue t = 24hrs G = 15.2 noc 0.2ug/ml	[04_24_03.cel] [annot]	[04_24_03.exp]		
042404	Heterozygous	04_01_1 het pool, DMSO/glycerol issue t = 24hrs G = 14.9 ben 0.4ug/ml	[04_24_04.cel] [annot]	[04_24_04.exp]		
050101	Heterozygous	04_25_01 het -70C	[05_01_01.cel] [annot]	[05_01_01.exp]		
050102	Heterozygous	04_25_01 t = 0 (on) G = 10	[05_01_02.cel] [annot]	[05_01_02.exp]		
050103	Heterozygous	04_01 het noc 4ug/ml 24hrs	[05_01_03.cel] [annot]	[05_01_03.exp]		
050802	Heterozygous	het 04_25_01 noc 2ug/ml 24hr G = 17	[05_08_02.cel] [annot]	[05_08_02.exp]		

Automated Statistical Analysis/QC



Easy access to past analyses



Analysis	Chip1	Chip2
041702 vs.041703	[04_17_02.cel]	[04_17_02.exp]
041702 vs.042401	[04_17_02.cel]	[04_17_02.exp]
041702 vs.042403	[04_17_02.cel]	[04_17_02.exp]
041702 vs.050103	[04_17_02.cel]	[04_17_02.exp]
042401 vs.050103	[04_24_01.cel]	[04_24_01.exp]
050101 vs.050102	[05_01_01.cel]	[05_01_01.exp]
050102 vs.050802	[05_01_02.cel]	[05_01_02.exp]
050902 vs.051105	[05_09_02.cel]	[05_09_02.exp]
051101 vs.051102	[05_11_01.cel]	[05_11_01.exp]
051101 vs.051103	[05_11_01.cel]	[05_11_01.exp]
051603 vs.051604	[05_16_03.cel]	[05_16_03.exp]
051603 vs.052202	[05_16_03.cel]	[05_16_03.exp]
052203 vs.052204	[05_22_03.cel]	[05_22_03.exp]
052205 vs.052206	[05_22_05.cel]	[05_22_05.exp]
05_15_01 vs.05_15_02	[05_15_01.CEL]	[05_15_02.CEL]
05_15_08 vs.05_15_07	[05_15_08.CEL]	[05_15_07.CEL]
Adam1 vs.Adam2	[080101.CEL]	[080102.CEL]
Adam1 vs.Adam3	[080101.CEL]	[080901.CEL]
Adam1 vs.Fake	[080101.CEL]	[080101_s.CEL]
Adam1 vs.Test	[080101.CEL]	[TPZ072600.CEL]
Adam2 vs.Fake	[080102.CEL]	[080101_s.CEL]
Adam2 vs.Test2	[080102.CEL]	[TPZ0815_3.CEL]
Adam3 vs.Adam4	[080901.CEL]	[081100.CEL]
Guri1 vs.Guri2	[04_17_01.cel]	[04_17_01.exp]

Automated Analysis/Target Hyp.

QuickRat: Chip Ratio Web - Microsoft Internet Explorer

Address http://genomics.lbl.gov/~aparkin/Projects/Giaever/DTrials/Experiments/QuickRat/cgi-bin/ShowAnalysis.cgi?CMP=05_15_08.vs.05_15_07

Bio/Spice::QuickRat

HHMI
University of California
Lawrence Berkeley National Laboratory

Analysis Pages

05_15_08	05_15_07
Chip 1 Avg= 7451.722395	Chip 2 Avg= 6998.749015
CHIP Comparison	Quality vs. Ratio I2/I1
[data]	[data]
Histogram: Avg. Strain Intensity	Histogram: (I2/I1)*(Avg1/Avg2)

QuickRat: Chip Ratio Web - Microsoft Internet Explorer

[data] Histogram: Avg. Strain Intensity

[data] Histogram: (I2/I1)*(Avg1/Avg2)

Avg= 1.012888 Std=0.68

QuickRat: Chip Ratio Web - Microsoft Internet Explorer

REGEXP: | Brightness Range: L H | Ratio Range: L H | Quality Range: L H

Sort By: Ratio Intensity Quality None | A D | Limit to results.

Query Results | Reset Form

QUERY RESULTS

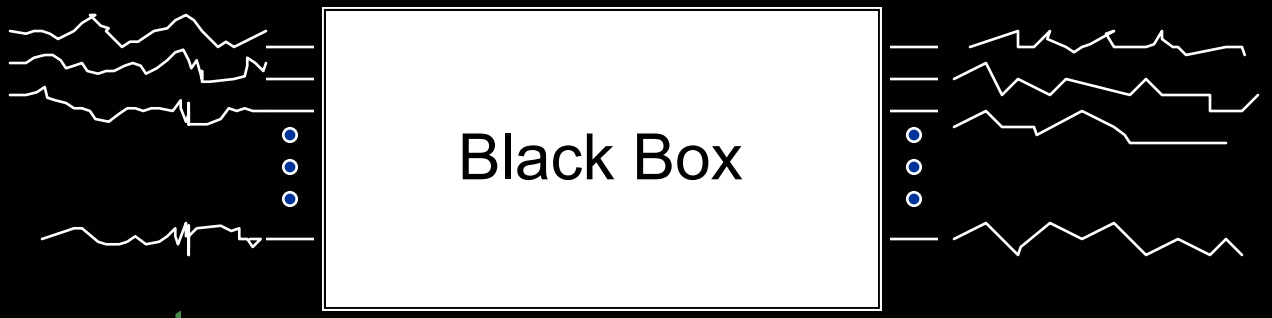
Strain	Gene			Viability
Ratio	Intensity	Quality	NTags	
YML124C	TUB3	tub3		INESSENTIAL
-1.211994	0.569563	8.213675	4	
alpha-3 tubulin				
alpha-tubulin				
YCR036W	RBK1	RBK1	RBK1	INESSENTIAL
-1.067519	0.344817	7.271917	4	
ribokinase				
ribokinase				
YHR156C				INESSENTIAL
-1.058958	0.538422	7.181411	4	
weak similarity to mouse kinesin KIF3B				
YCR028C	FEN2	FEN2	FEN2	INESSENTIAL
-1.032899	0.652064	6.990636	4	
similarity to allantoate permease transporter				
Amino acid permease				
YCR085W				INESSENTIAL
-1.031515	0.622995	6.984592	4	
hypothetical protein				
YGR029W	ERV1	ERV1	ERV1	ESSENTIAL
-0.985959	0.873954	6.654740	4	
mitochondrial biogenesis and regulation of cell cycle				

Network Deduction and Modules

An incredibly naïve approach

Correlation Metric Construction

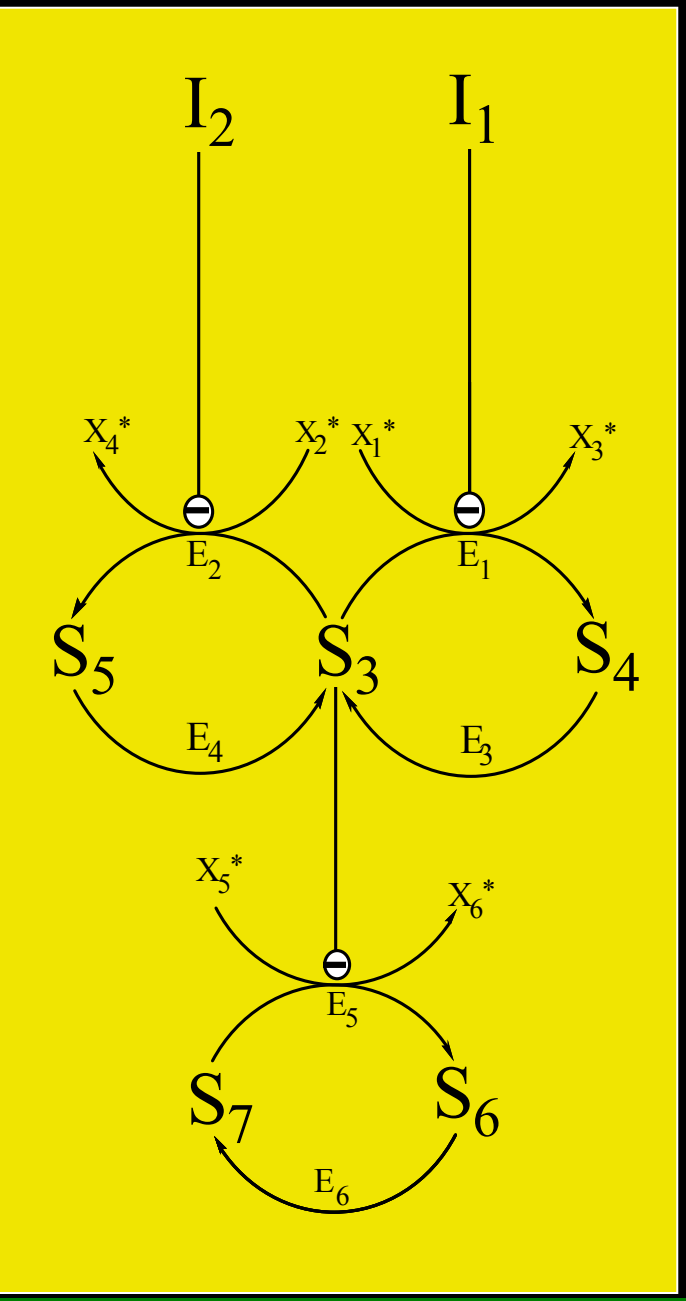
A method to deduce reaction pathways directly from concentration time-series measurements



Temperature
pH
Pharmaceutical levels
etc.

Concentrations
Transcript Levels
etc.

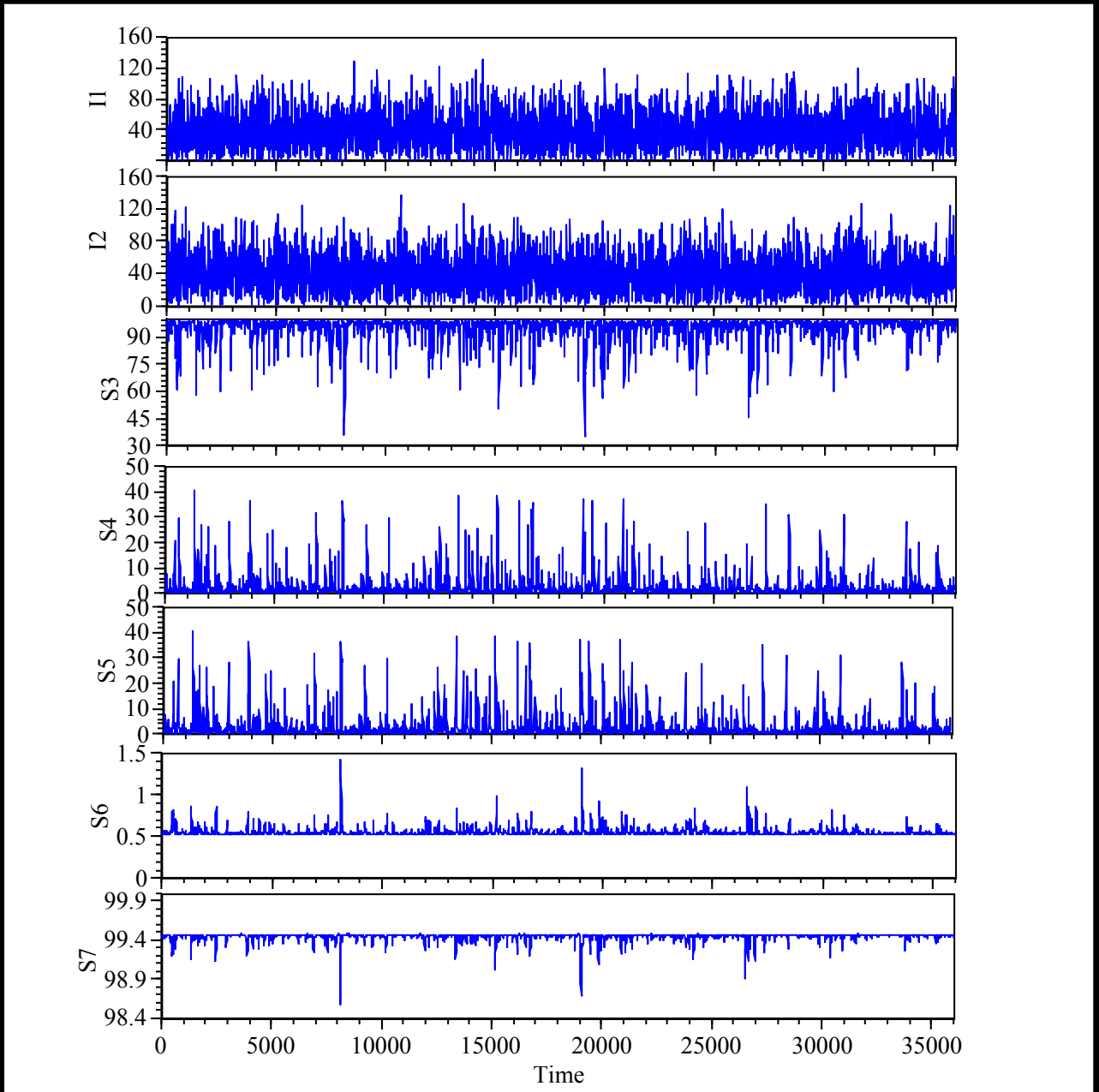
Pei-dong Shen, Michael Samoilov, John Ross



This is an abstract biochemical NAND gate

Based on a mechanism at the heart of switching between the glycolytic and gluconeogenic modes of the hexose phosphate pathway.

Can we deduce the network structure by perturbing it with inflows of the inputs and measuring the response of the concentrations?


I1
160
120
80
40
I2
160
120
80
40
0
S3
90
75
60
45
30
S4
50
40
30
20
10
0
S5
50
40
30
20
10
0
S6
1.5
1
0.5
0
S7
99.9
99.4
98.9
98.4
0
5000
10000
15000
20000
25000
30000
35000
Time

Measures of Dependency

Linear Correlation

- Measures the linear relationship between variables.
- May be extended to multiple dependencies (i.e. y= f(x1,x2,x3...) by assumption of a linear regression model.
- Very difficult to tell the significance of a given correlation since no distributional assumptions are made.

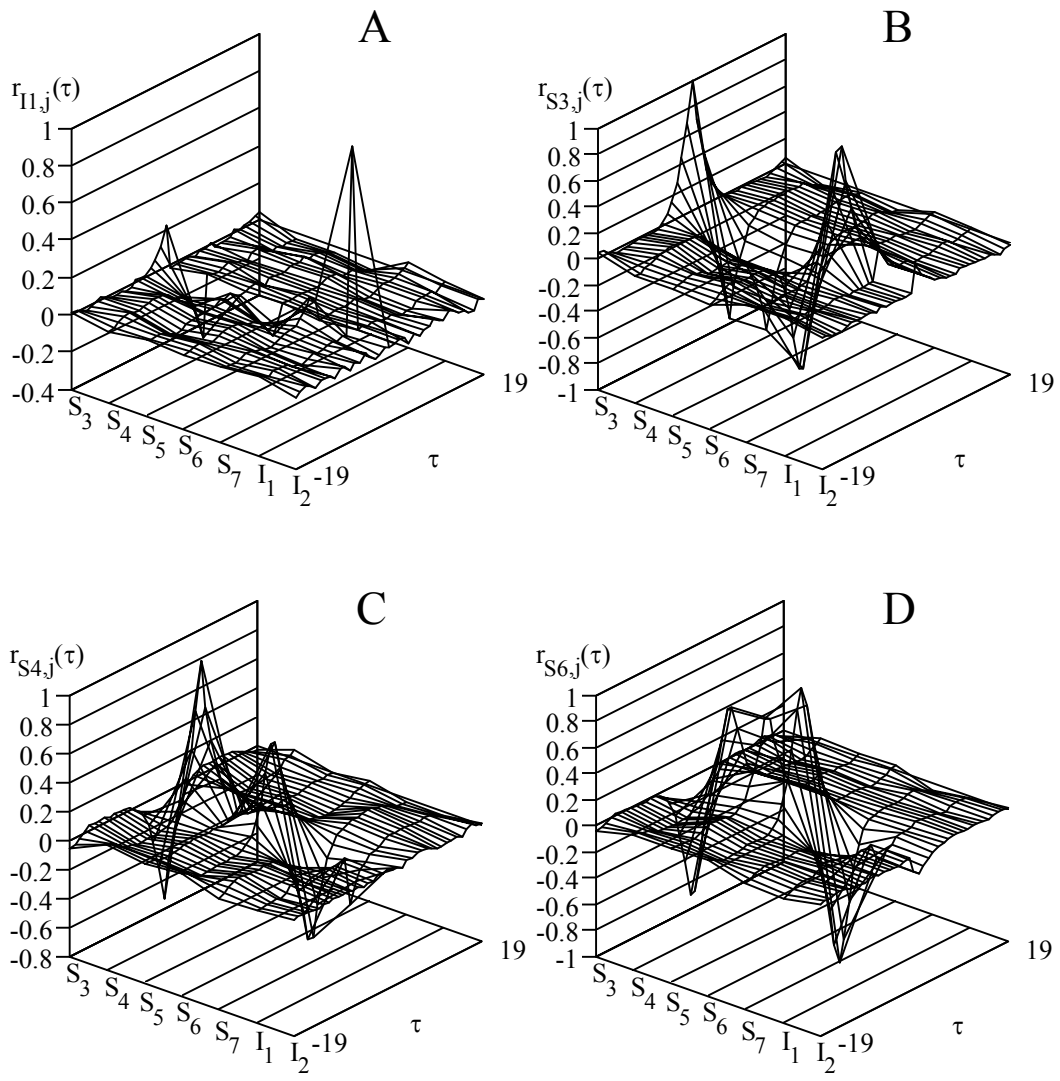
Non-Parametric Rank Correlation (Spearman)

- Measures monotonic relationships between variables.
- Like Linear Correlation except that distribution of numbers is now known (uniform, exactly).
- Robust to data defects.
- Significances may be calculated...weak dependencies may be missed.

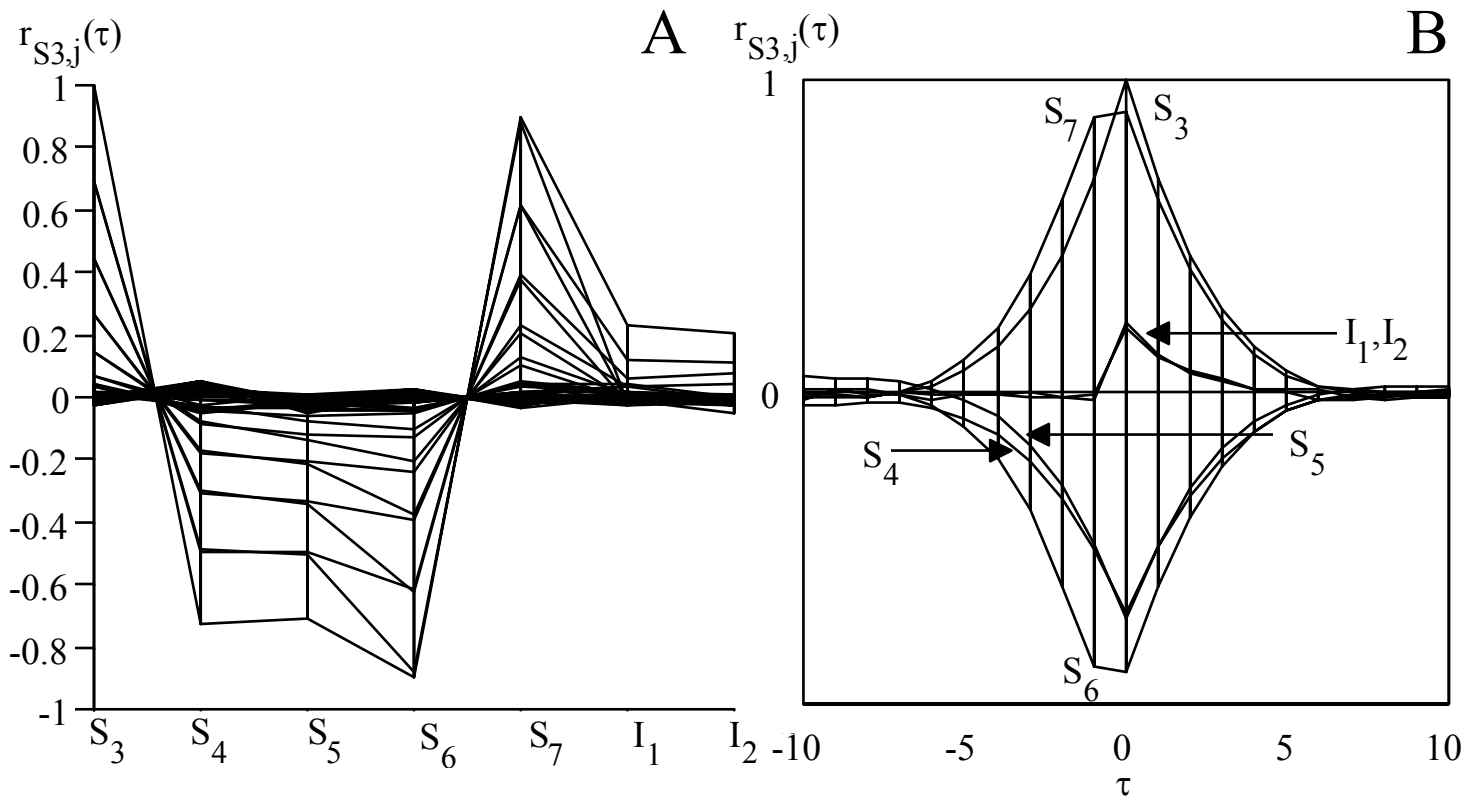
Transinformation

- Measures the constraint on one variable given knowledge of another; i.e., requires that one variable merely be a function of the other.
- The distribution is the quantity actually calculated. c2 then provides an accurate measure of significance.
- Multiple dependencies easily incorporated by increasing the dimension of the distribution.

Chemical NAND Gate Correlation Functions



Correlation Function Projection



Multidimension Scaling Solutions

A) Eigenvectors, $\mathbf{z}_k$:

point/z_k	1	2	3	4	5	6	7
1 (I_1)	6.68e-01	-5.84e-01	4.05e-01	5.51e-02	1.77e-02	-8.20e-09	-3.52e-10
2 (I_2)	7.00e-01	5.26e-01	-4.30e-01	4.93e-02	1.42e-02	-8.20e-09	-3.52e-10
3 (S_7)	-4.20e-01	7.29e-03	-8.16e-03	2.05e-01	1.90e-03	-6.82e-09	-7.67e-09
4 (S_6)	-4.20e-01	7.29e-03	-8.16e-03	2.05e-01	1.90e-03	-9.58e-09	6.97e-09
5 (S_4)	-1.44e-01	-5.51e-01	-4.02e-01	-1.60e-01	-7.55e-02	-8.20e-09	-3.52e-10
6 (S_5)	-7.15e-02	5.60e-01	4.30e-01	-1.38e-01	-7.65e-02	-8.20e-09	-3.52e-10
7 (S_3)	-3.14e-01	3.49e-02	1.27e-02	-2.16e-01	1.16e-01	-8.20e-09	-3.52e-10

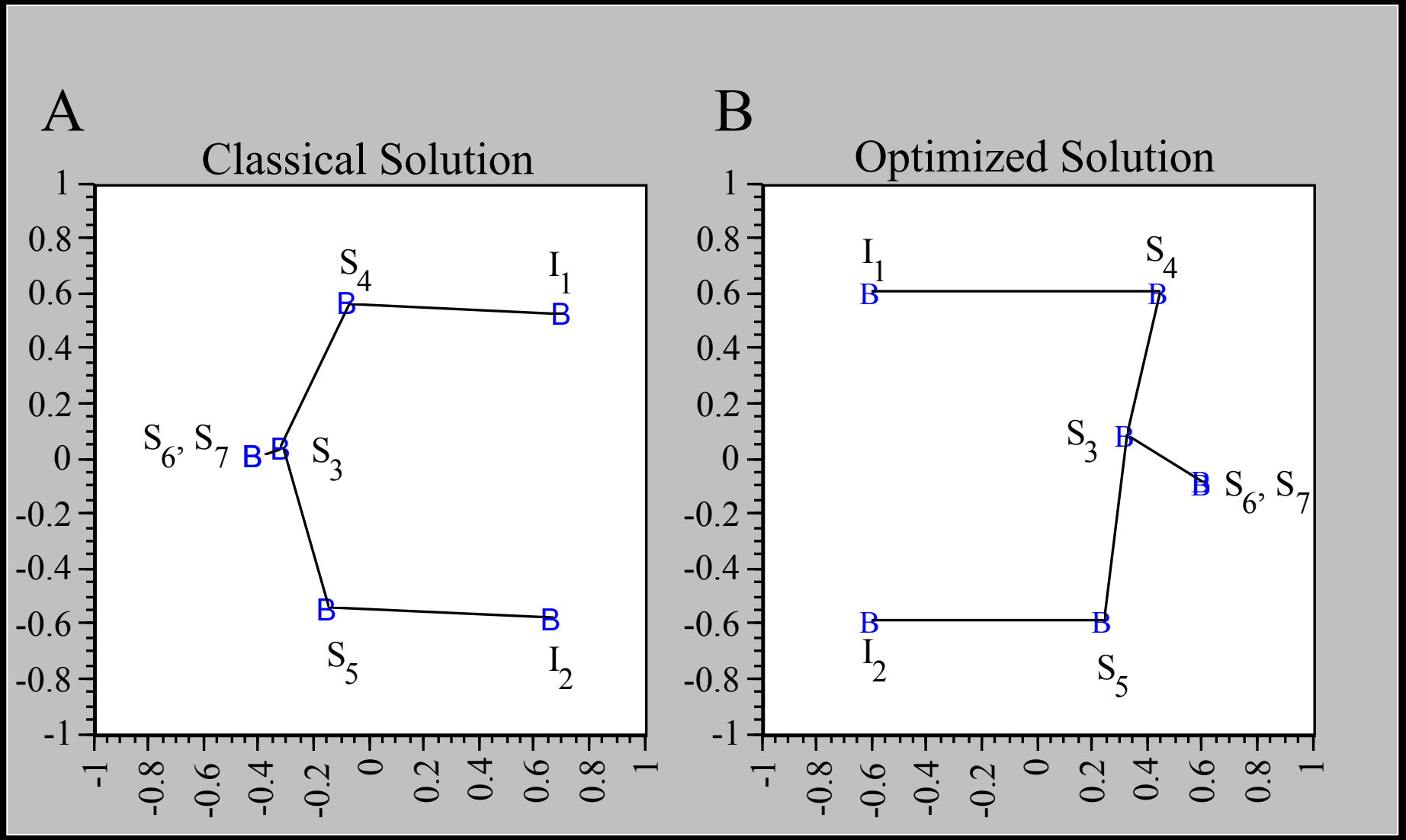
B) Eigenvalues, l_k:

k	l_k	$a_{1,k}$	$a_{2,k}$
1	1.413496e+00	3.978311e-01	4.937104e-01
2	1.237497e+00	7.461268e-01	8.721279e-01
3	6.958617e-01	9.419783e-01	9.917824e-01
4	1.805556e-01	9.927960e-01	9.998381e-01
5	2.559576e-02	1.000000e+00	1.000000e+00
6	4.747935e-16	1.000000e+00	1.000000e+00
7	1.080736e-16	1.000000e+00	1.000000e+00

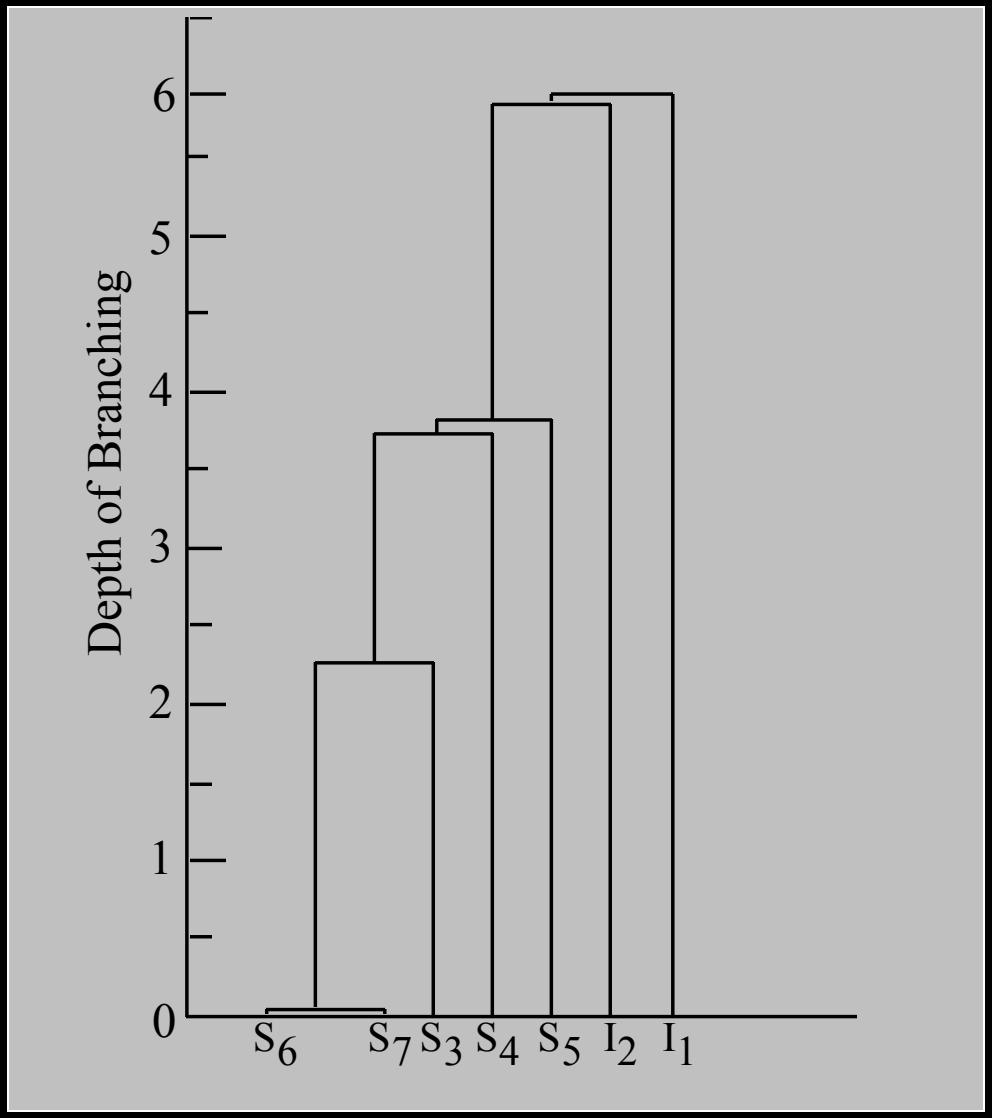
C) Significant Connections

	S_4	S_5	S_6	S_7
I_1	-0.31			
I_2		-0.32		
S_3	-0.72	-0.71	-0.90	0.90
S_6				-1.00

MDS Solutions: Projection and Squash



Hierarchical Cluster Diagram



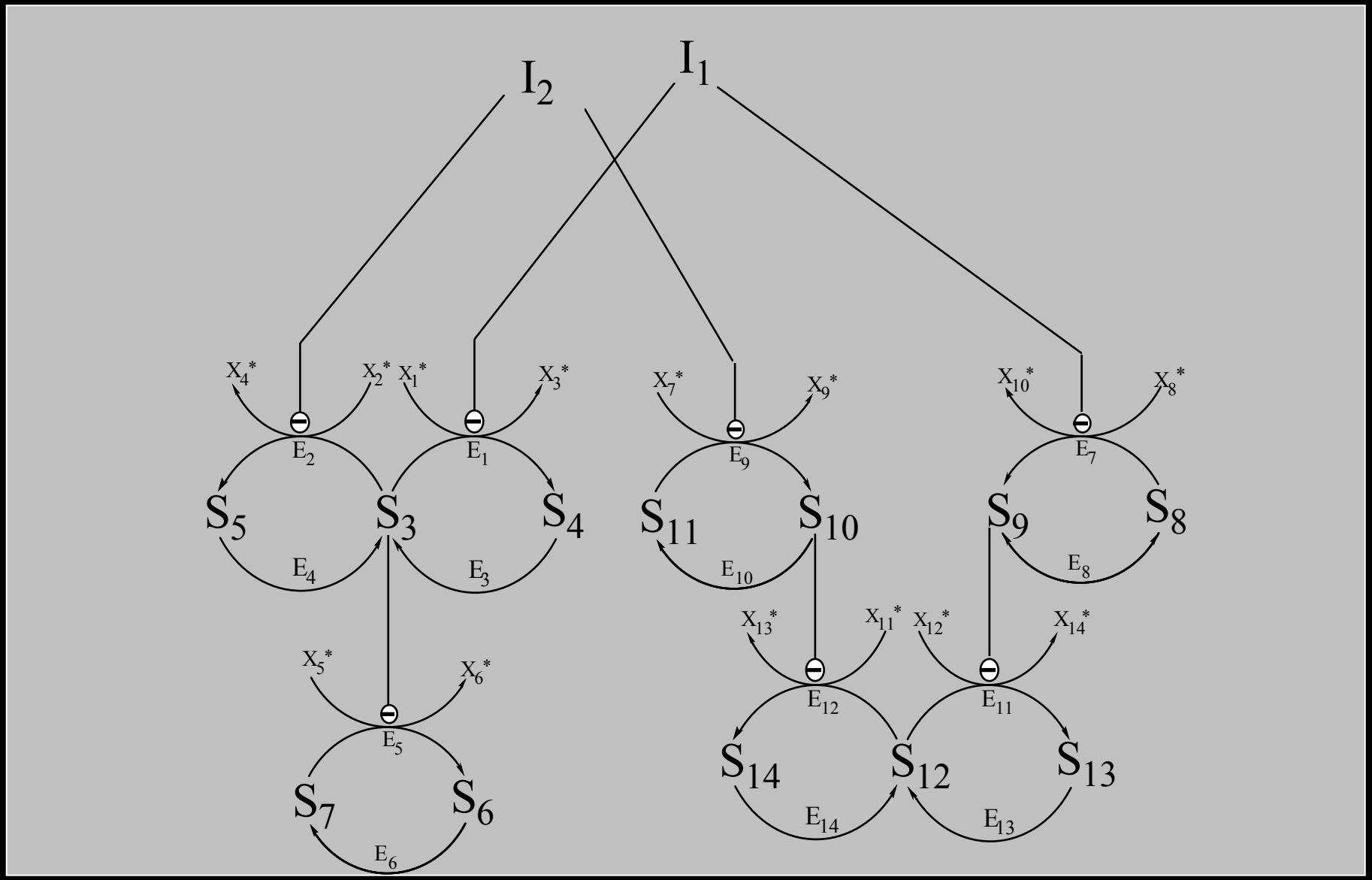
Tightly coupled species are grouped together.

Methodology for determining tightly coupled pathways?

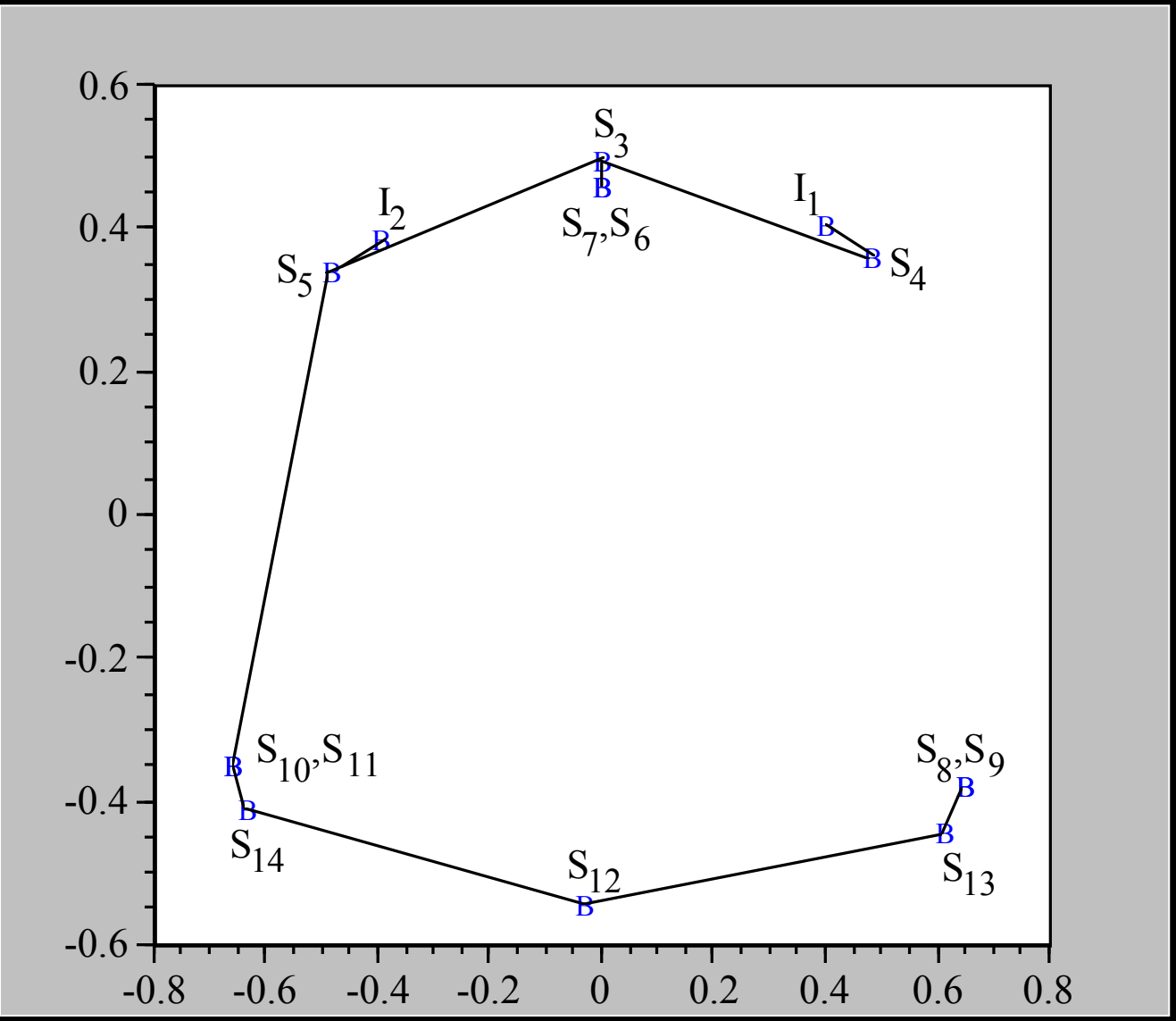
Or chemical subsystems that may be analyzed outside of the rest of the circuit?

MORE ON MODULARITY: STAY TUNED!

A more complex case



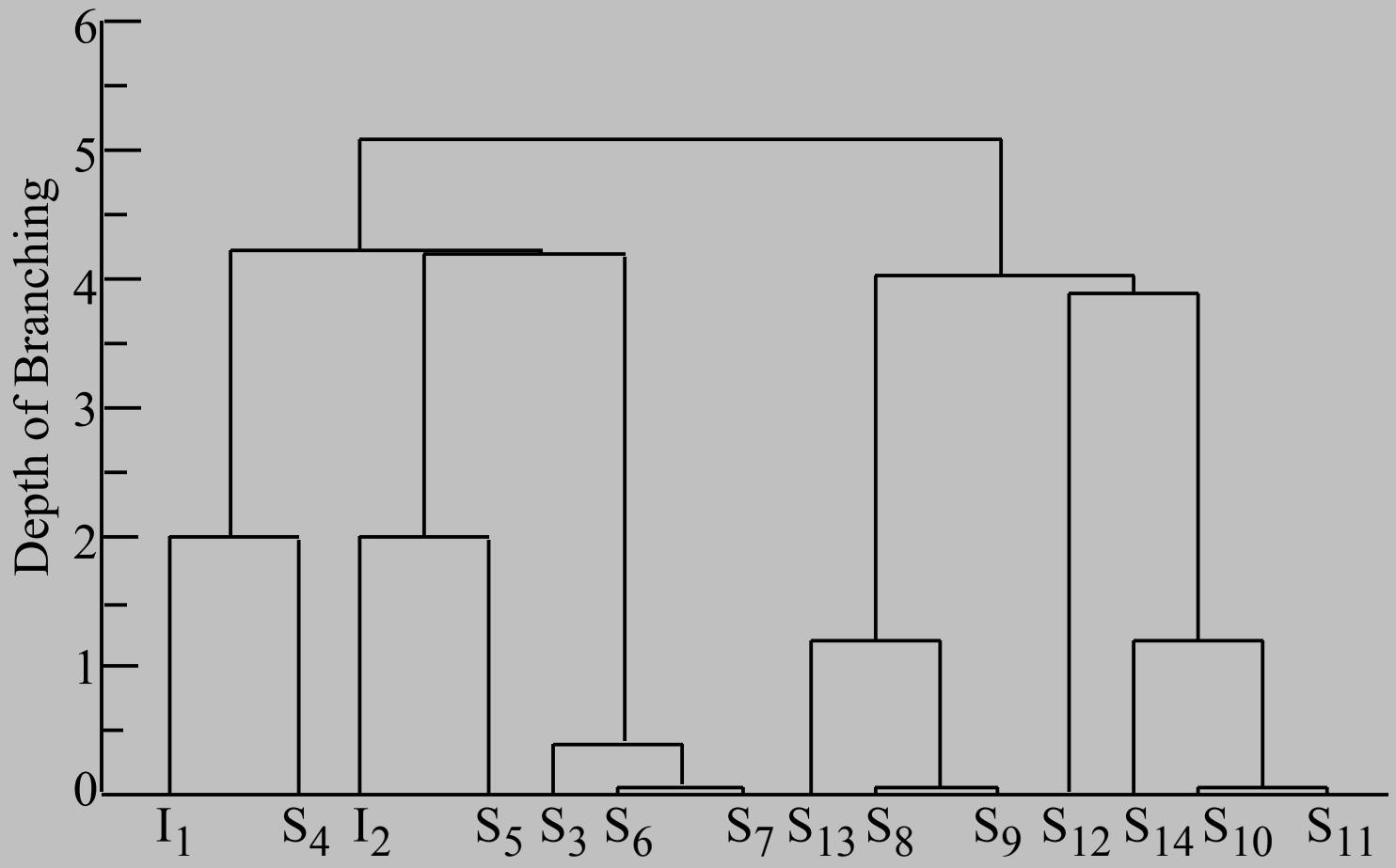
The MDS Diagram



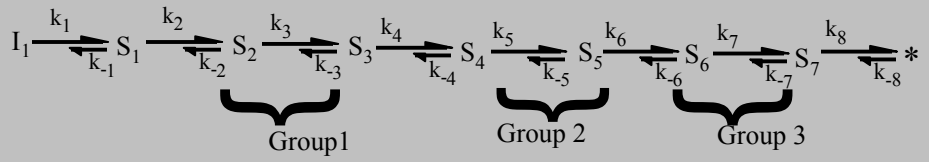
Note that the two parallel pathways groups on different sides of the diagram

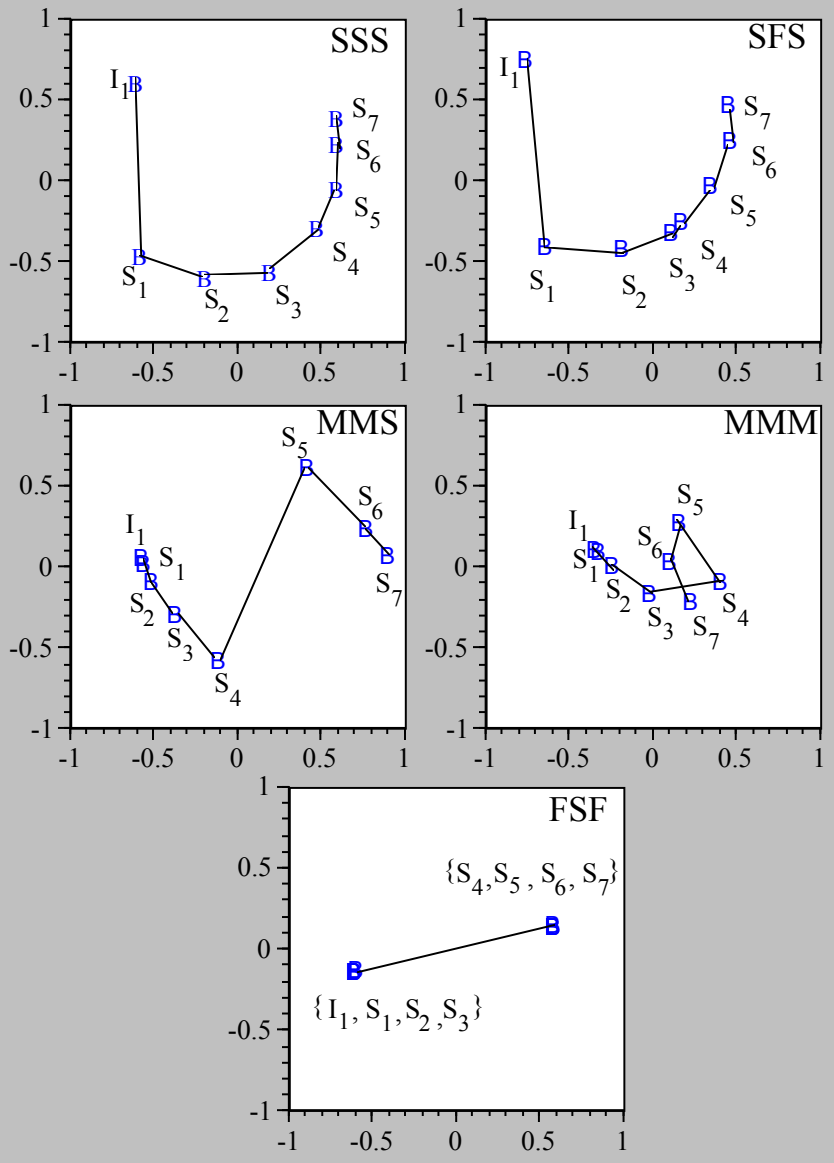
But also note the spurious connection between S5 and S10,11.

Hierarchical Cluster

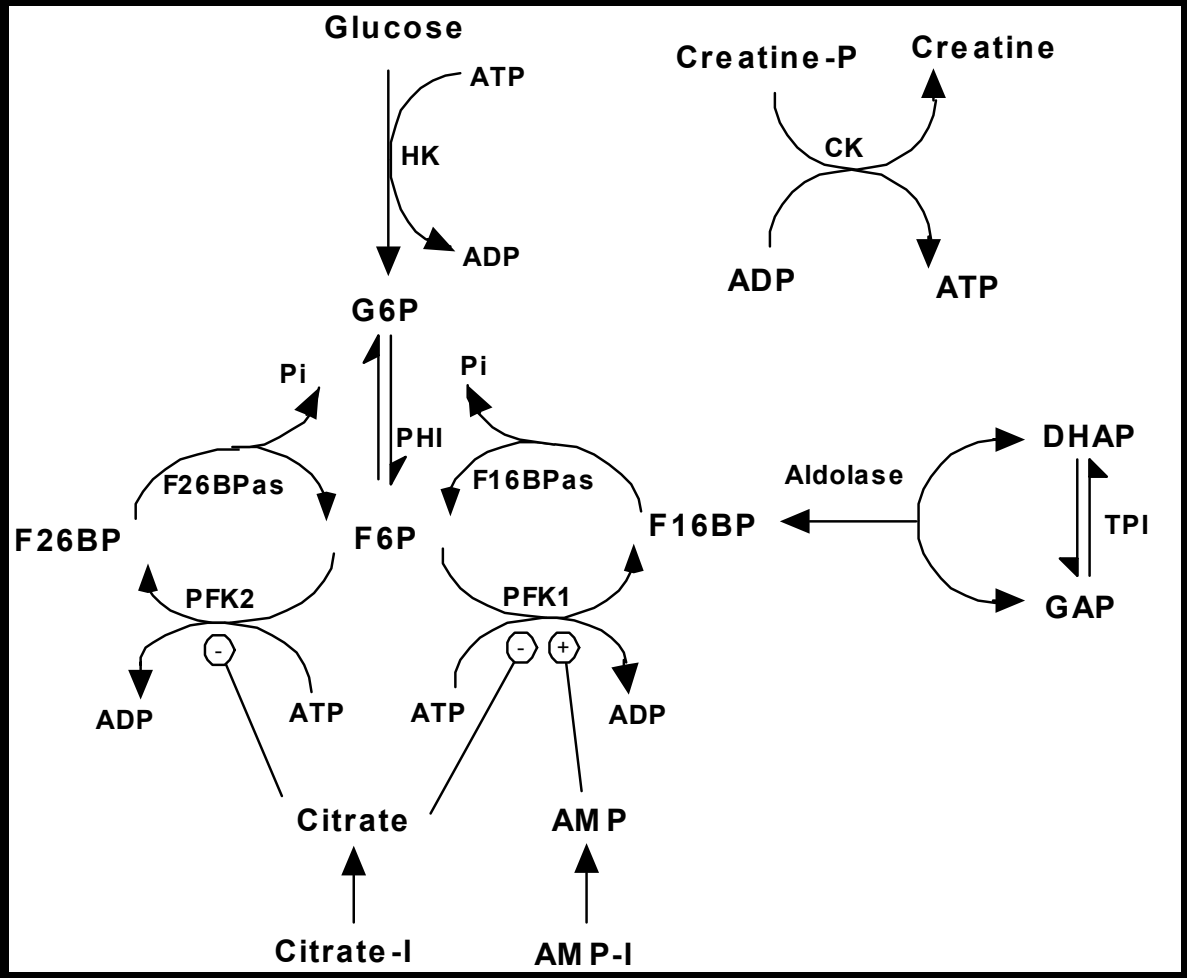


Reaction Chain Cases

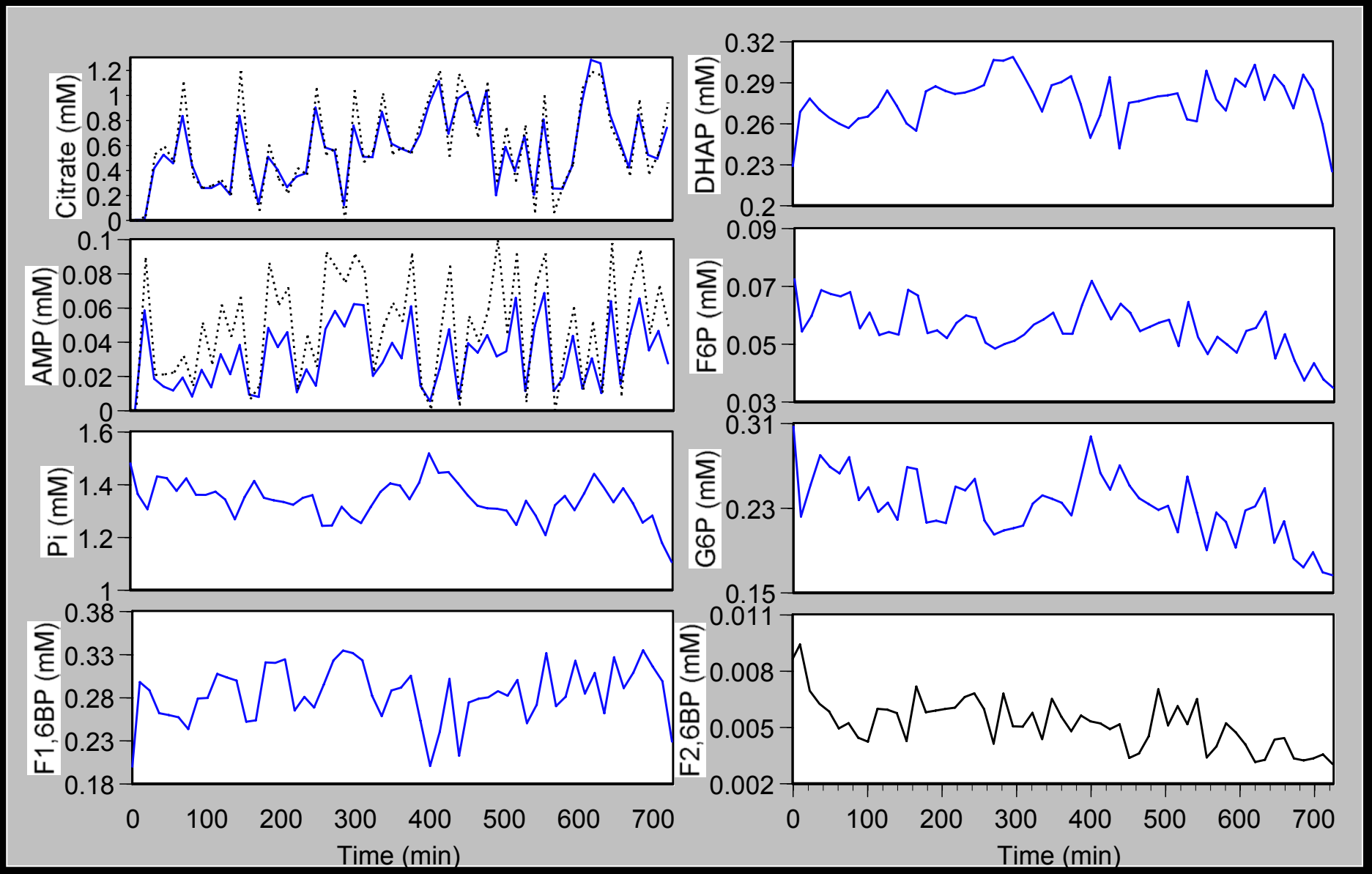


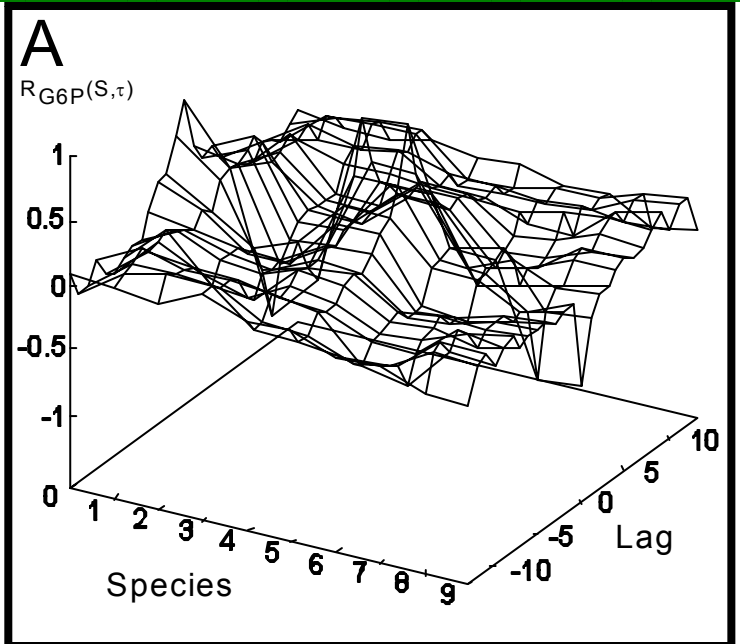


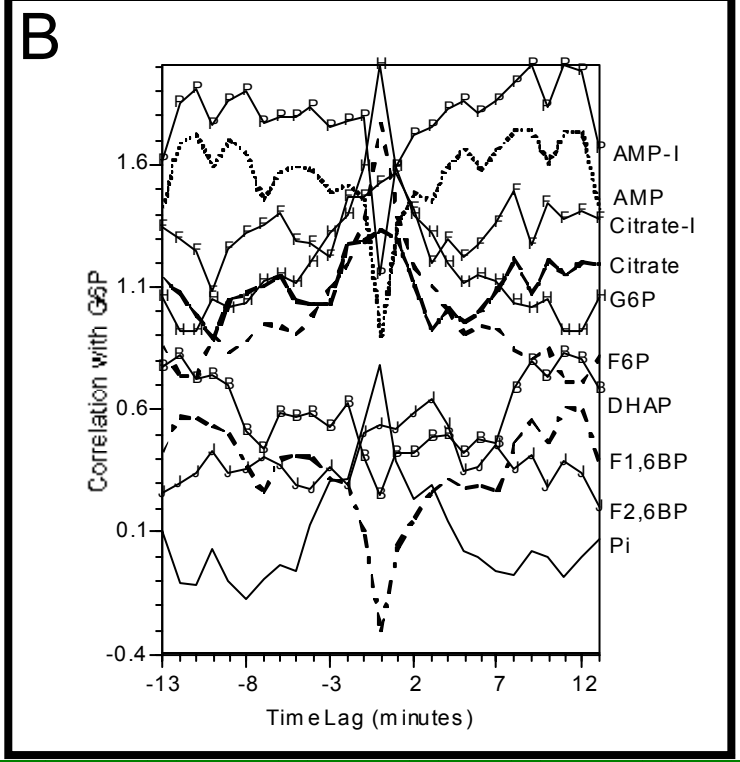
Experimental Test System



Capillary Electrophoresis Data



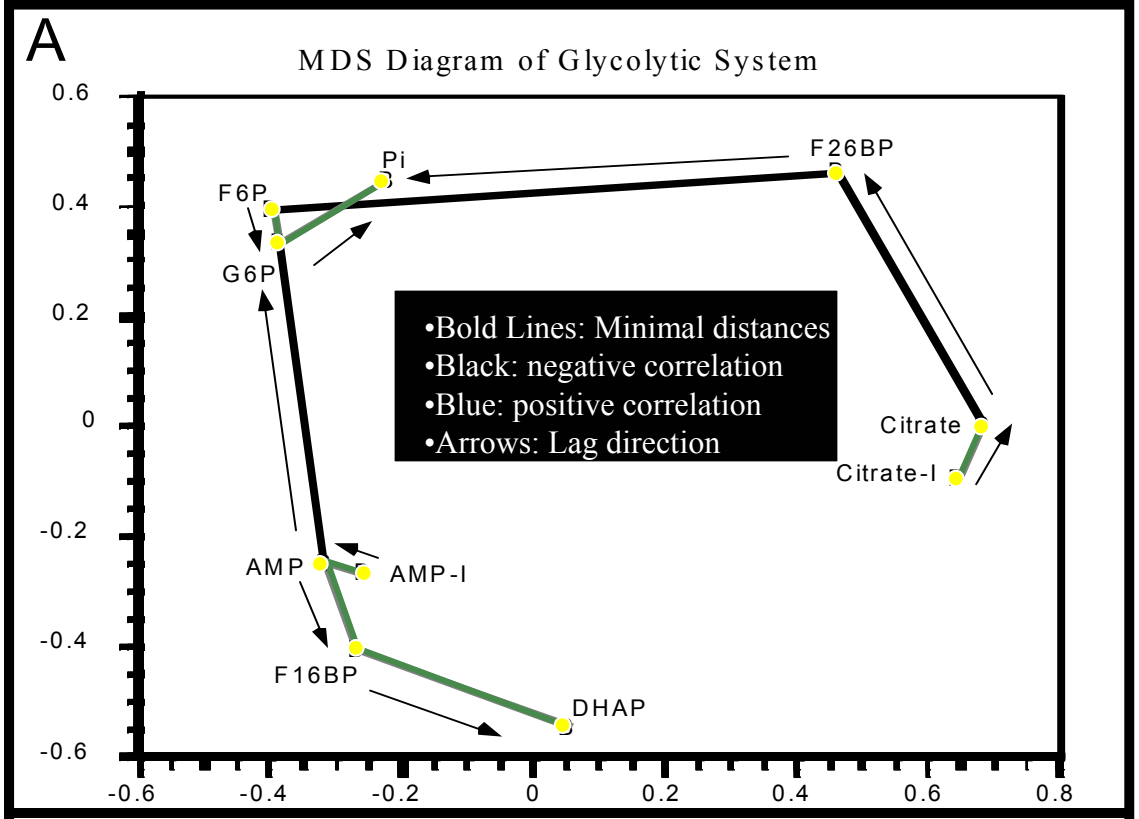


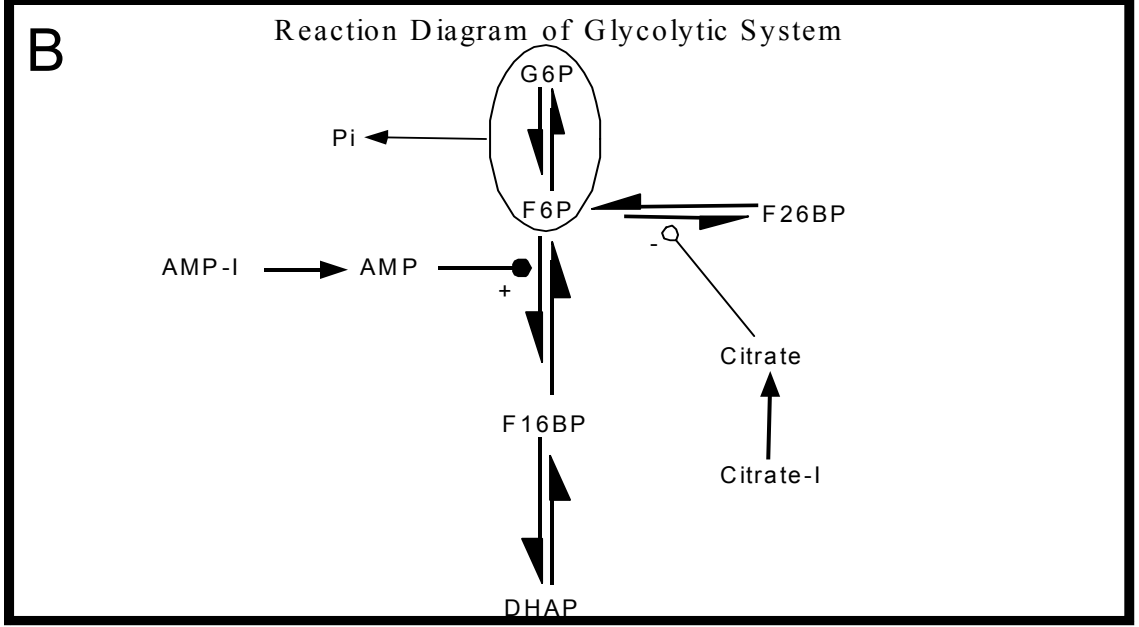


Experimental correlation of all 10 measured glycolytic intermediates with G6P.

Both strength of interaction and temporal ordering are implicit in time-dependent correlation function

Projection of the surface down the time-lag axis. The results of each species are offset a bit from zero.





Multidimensional Scaling (MDS) diagrammatic summary of correlation functions.

Diagram summarizes:

- Interaction strengths
- Temporal ordering
- Probable network structure

"Expert system" prediction of reaction pathway from correlation matrix/MDS analysis

Correlation misses relationships such as this. Information is a better metric.

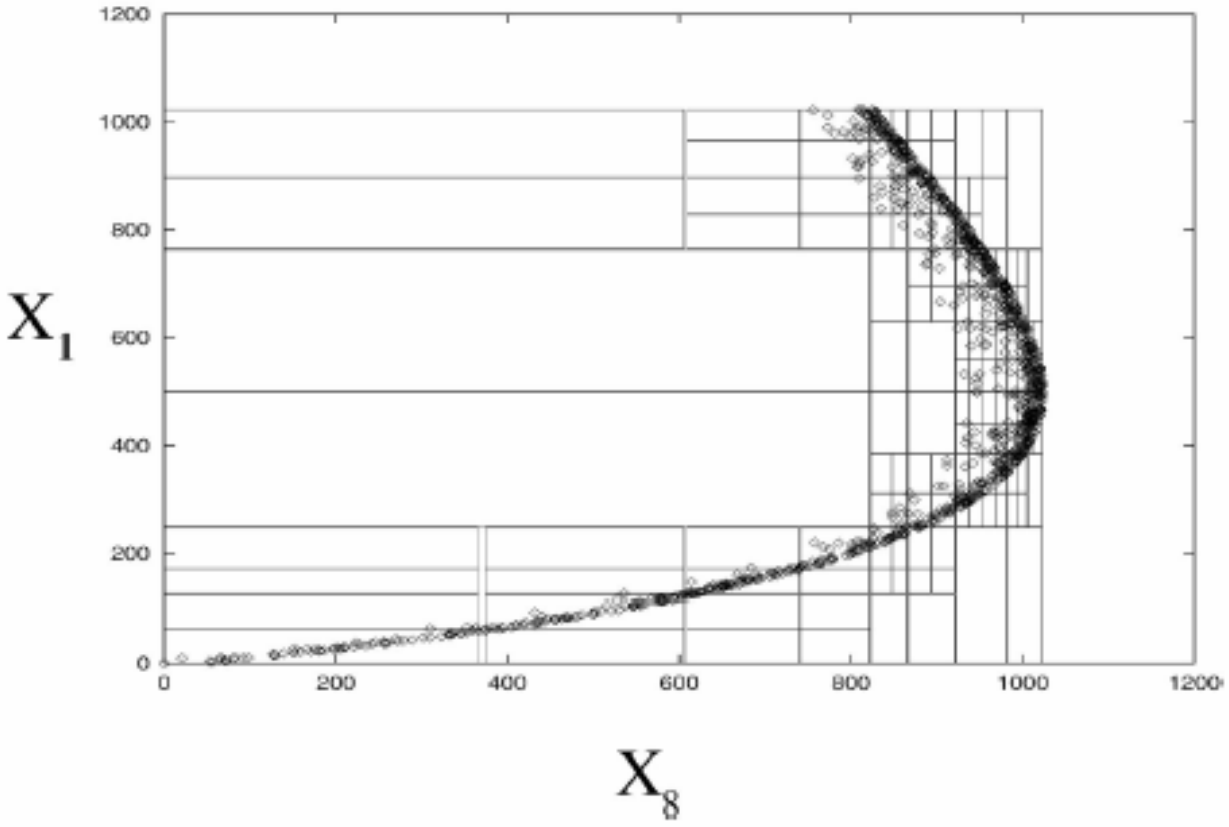


FIG. 5. The diamonds plot the values of X_1 vs X_8 obtained from the simulated time series. The rectangles are the result of a partioning algorithm, see the text. From Ref. 32.

Samoilov, Arkin, Ross, Chaos, 11(1):108-114

Data problems: Theoretically $N= 5*Q^M$

Some short thoughts on modules

Considerations

How do we define modules?

1. Repeated units
2. Evolutionary conservation
3. Time scale separation
    a. fast and slow manifolds (Michaelis Menten)
    b. on at different times
4. Spatial/Structural separation
    a. no links between two subsystems
    b. separated in space literally
    c. weak coupling
5. Individually controllable
    a. high impedances between subsystems?

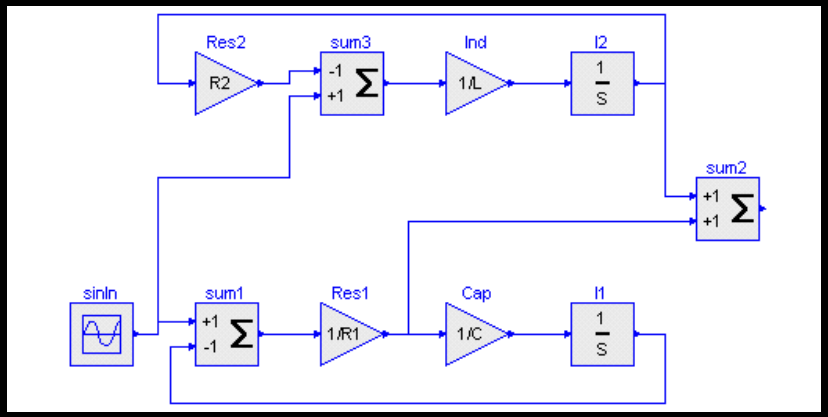
Elliott will define more!

Schematics for cells: Modules

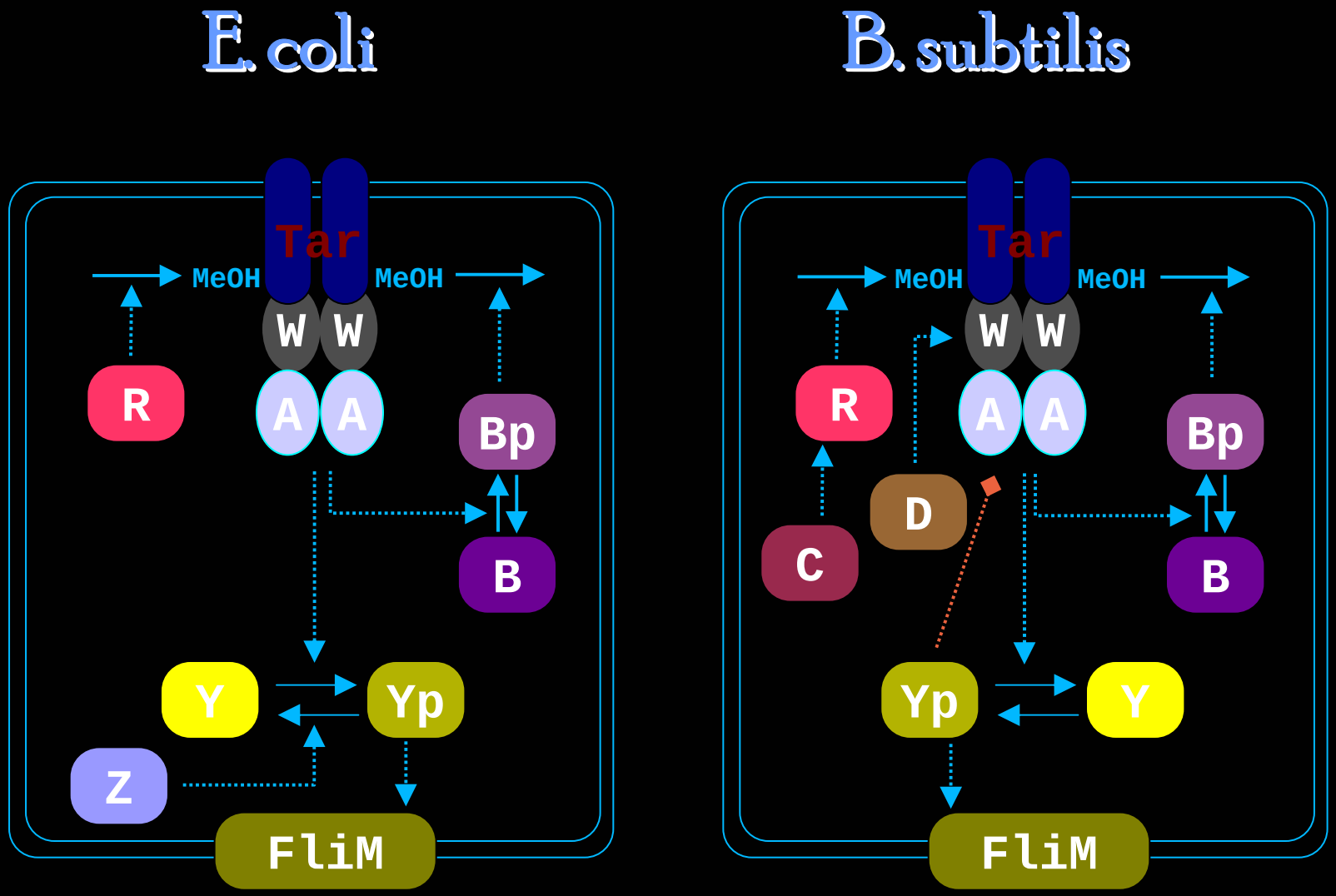
Motif= A repeated pattern of interactions among objects

Module

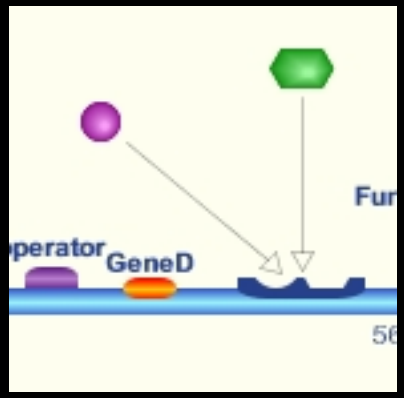
1. A motif
2. An elementary functional unit
3. A compound functional unit that may be abstracted

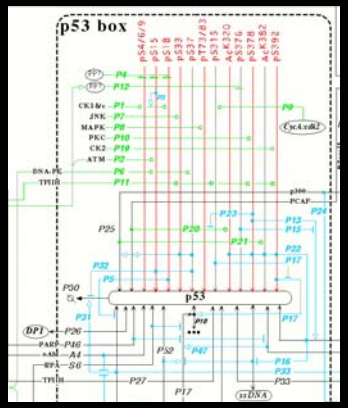


Chemotaxis Signal Transduction Pathways. A motif? A module?



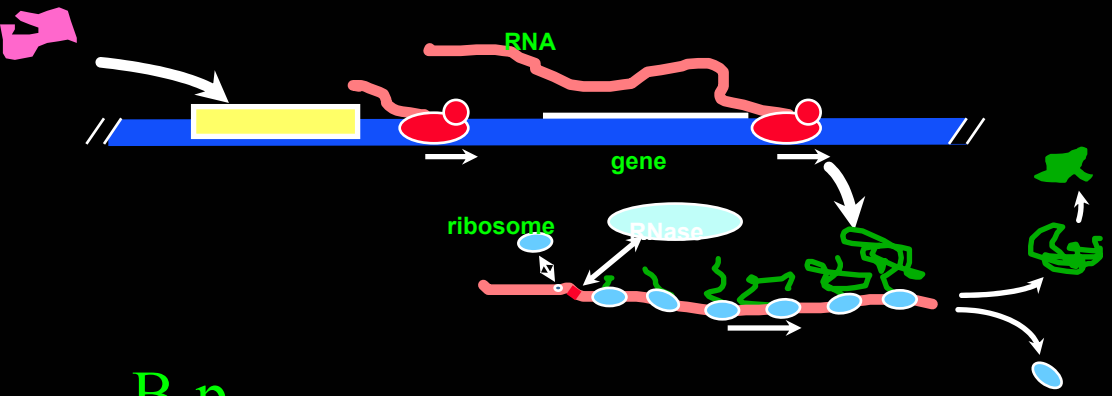
Representation problems

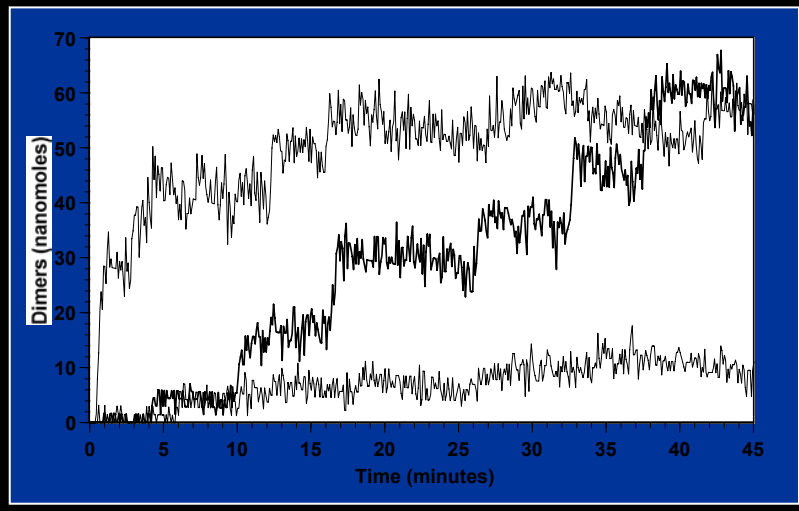


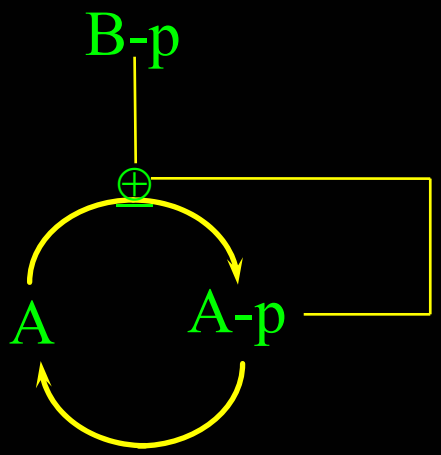


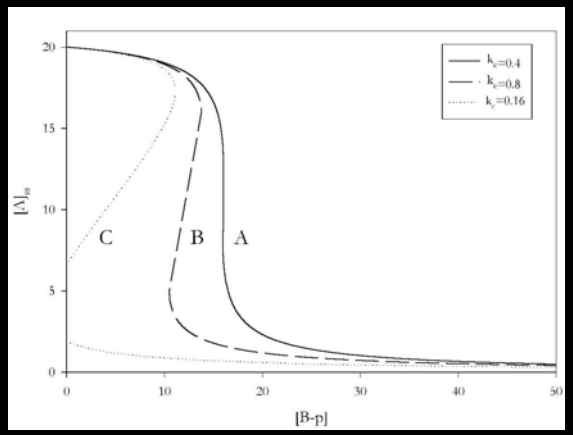
Problems of combinatorics

Problems of physics



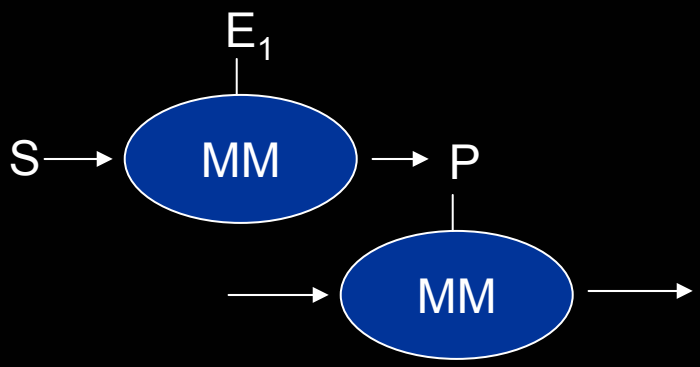






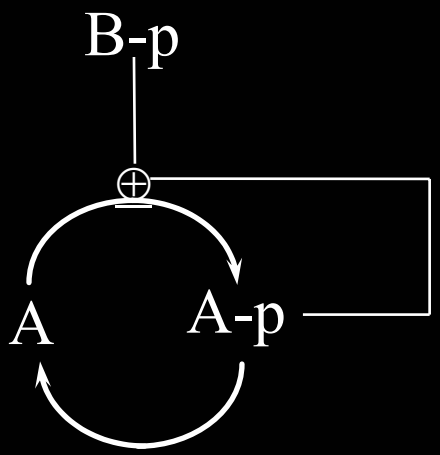
Problems of parameter sensitivity

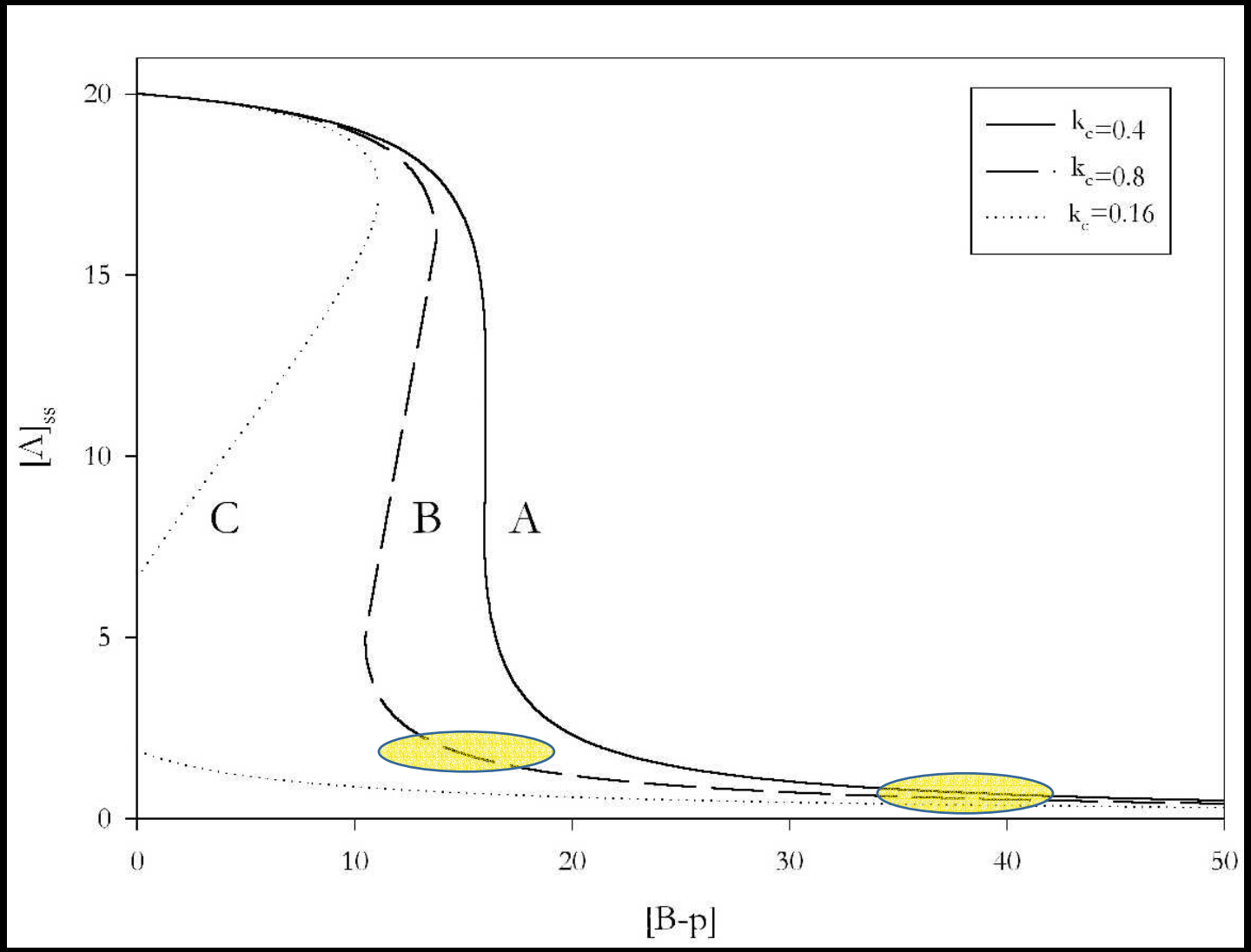
Problems of abstraction



Problems of assumptions

Problems of composition





Brief Digression: Chemical Impedance

I→A→*

$$\frac{dA}{dt} = k_1[I] - k_2[A] \quad \Rightarrow \quad A_{t\to\infty}(I) = \frac{k_1}{k_2}[I]$$

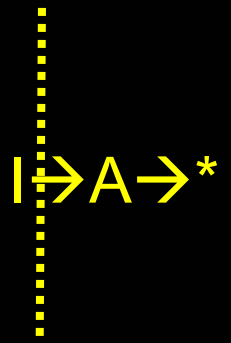
So A is the signal inside the cell that I is outside the cell.
What if A signals to downstream targets by reacting with them?

A+B*→C

$$\frac{dA}{dt} = k_1[I] - k_2[A] - k_3[A][B] \quad \Rightarrow \quad A_{t\to\infty}(I) = \frac{k_1[I]}{k_2 + k_3[B]}$$

The rates and concentrations of downstream processes degrade the signal from A.

Brief Digression: Chemical Impedance



$$\frac{dA}{dt} = k_1[I] - k_2[A] \qquad \Rightarrow \qquad A_{t\to\infty}(I) = \frac{k_1}{k_2}[I]$$

But what if reaction is by reversible binding?

$A + B^* \leftrightarrow C$

$$\frac{dA}{dt} = k_1[I] - k_2[A] - k_3[A][B] + k_4[C]$$

$$A_{t\to\infty}(I) = \frac{k_1[I]}{k_2}$$

The rates and concentrations of downstream processes don't affect the signal.