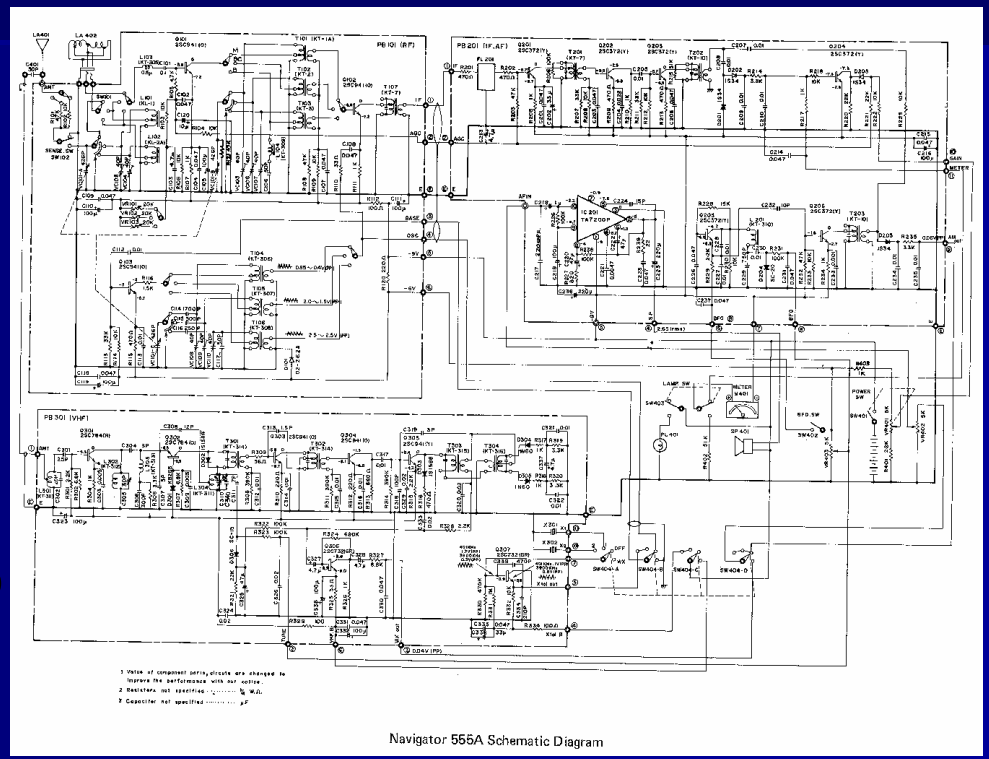
INSTITUTE FOR Systems Biology

Genetic Regulatory Network Modeling and Data Integration

ilya shmulevich

genetic networks

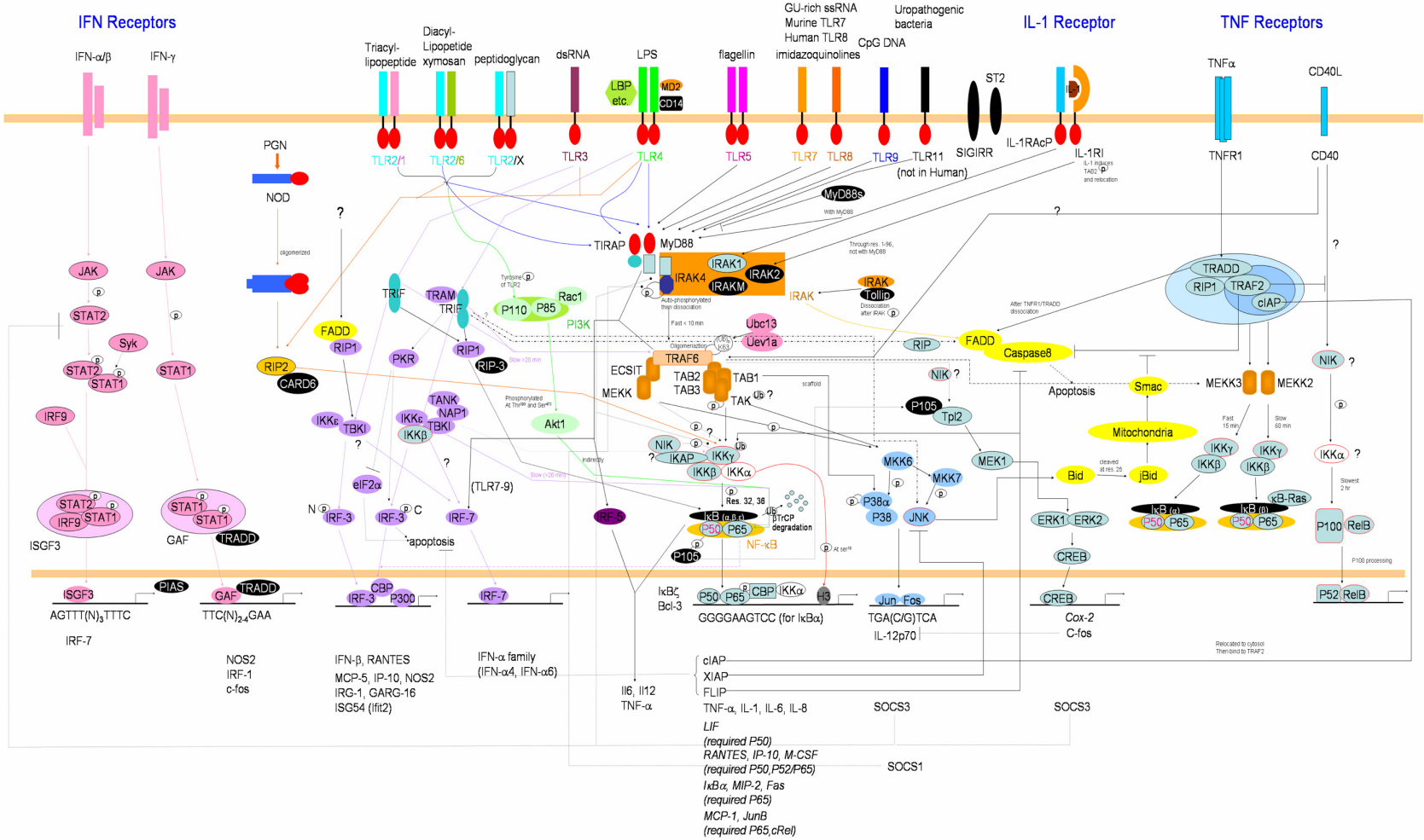
- Complex regulatory networks among genes and their products control cell behaviors such as:
  - cell cycle
  - apoptosis
  - cell differentiation
  - communication between cells in tissues
- A paramount problem is to **understand the dynamical interactions** among these genes, transcription factors, and signaling cascades, which govern the integrated behavior of the cell.



Navigator 555A Schematic Diagram

Analogy: circuit diagram

Map of the TLR signaling pathway in the macrophage

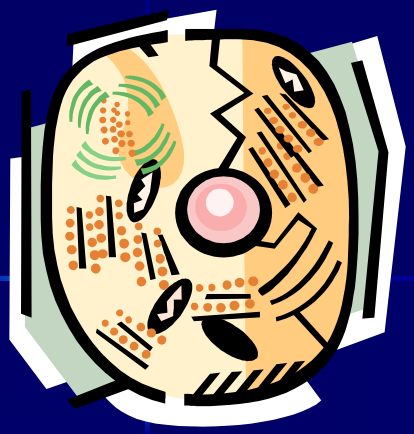


Goals

1. Discover and understand the underlying gene regulatory mechanisms by means of inferring them from data.
2. By using the inferred model, endeavor to make useful predictions by mathematical analysis and computer simulations.

Clinical Impact

- Model-based and computational analysis can
  - open up a window on the physiology of an organism and disease progression;
  - translate into accurate diagnosis, target identification, drug development, and treatment.



mRNA

Gene 1
Gene 2
Gene 3
Gene *n*

Microarray

Data

17.4
129
37
198.5
1020.4

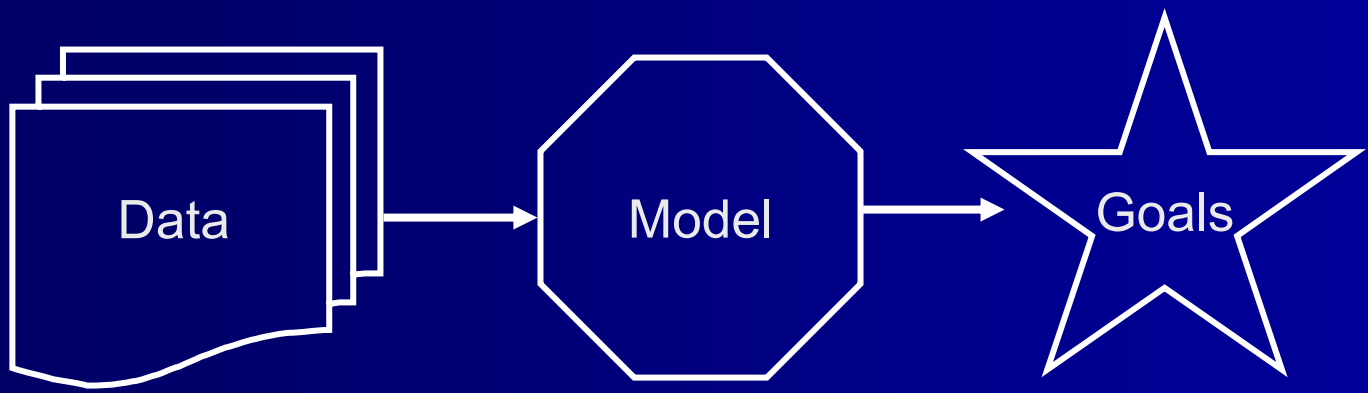
566.8
803.1

Extract Rules

A
B
C

What class of models should be chosen?

- The selection should be made in view of
  - data requirements
  - goals of modeling and analysis.



Classical tradeoff

- A "fine" model with many parameters
  - may be able to capture detailed "low-level" phenomena (protein concentrations, reaction kinetics);
  - requires large amounts of data for inference
- A "coarse" model with low complexity
  - may succeed in capturing only "high-level" phenomena (e.g. which genes are ON/OFF);
  - requires smaller amounts of data

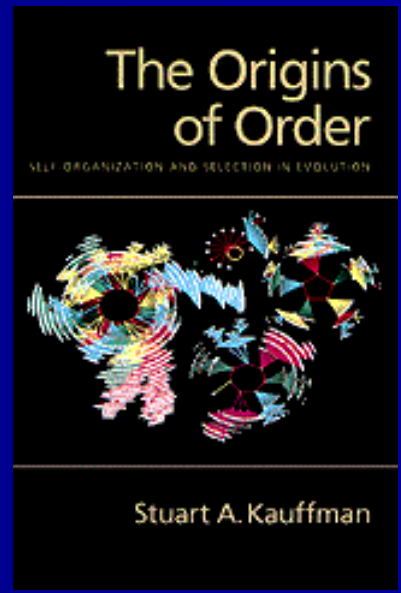
Ockham's Razor

- Underlies all scientific theory building.
- Model complexity should never be made higher than what is necessary to faithfully "explain the data."
- What kind of data do we have and how much?

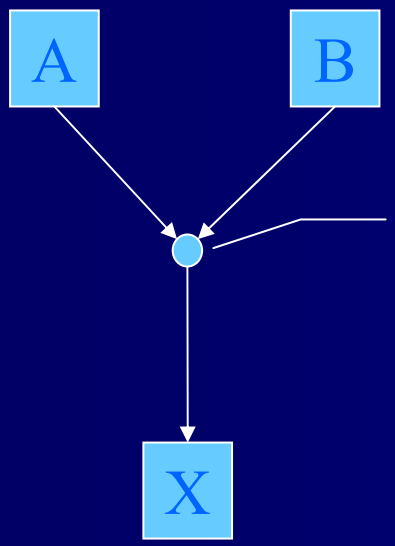
William of Ockham
(1280-1349)

Boolean Networks

1. To what extent do such models represent reality?
2. Do we have the "right" type of data to infer these models?
3. What do we hope to learn from them?



Basic Structure of Boolean Networks



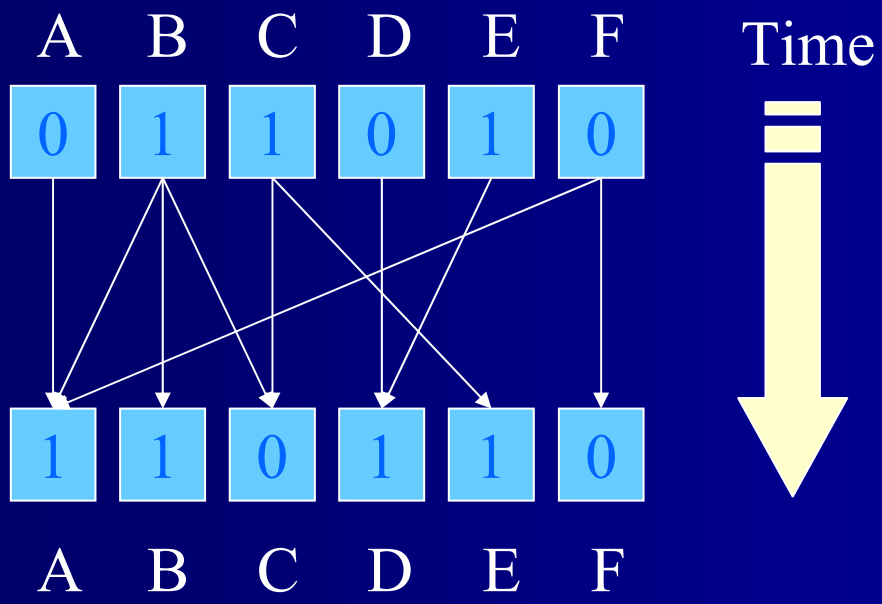
1 means *active/expressed*
0 means *inactive/unexpressed*

Boolean function

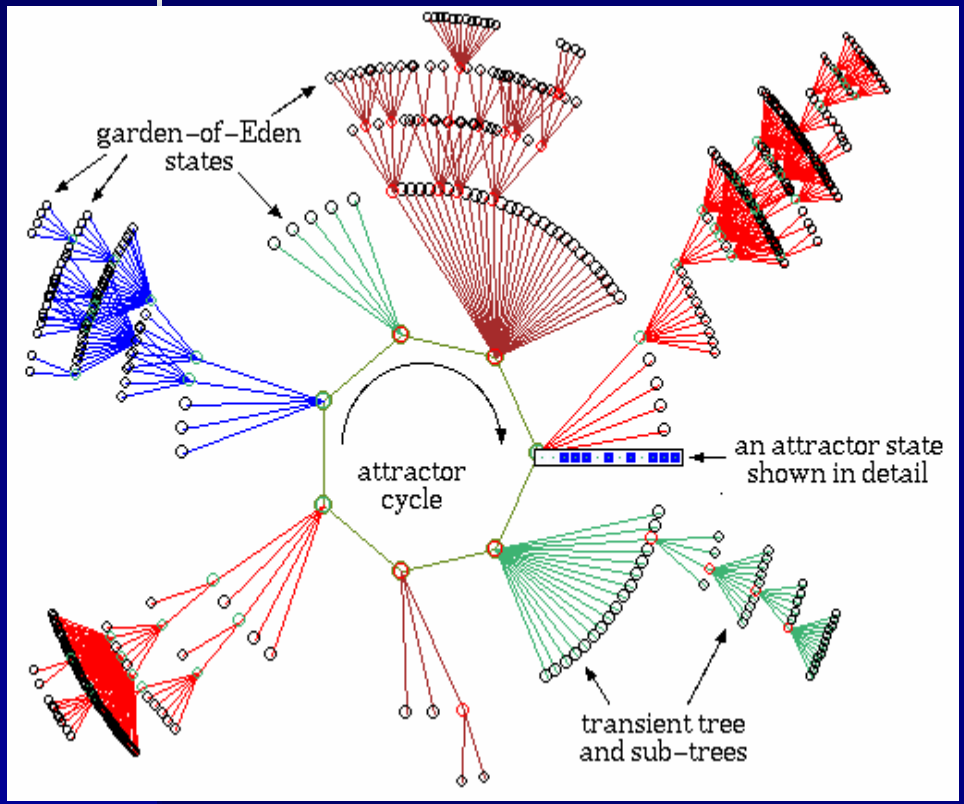
A	B	X
0	0	1
0	1	1
1	0	0
1	1	1

In this example, two genes (A and B) regulate gene X. In principle, any number of "input" genes are possible. Positive/negative feedback is also common (and necessary for homeostasis).

Dynamics of Boolean Networks



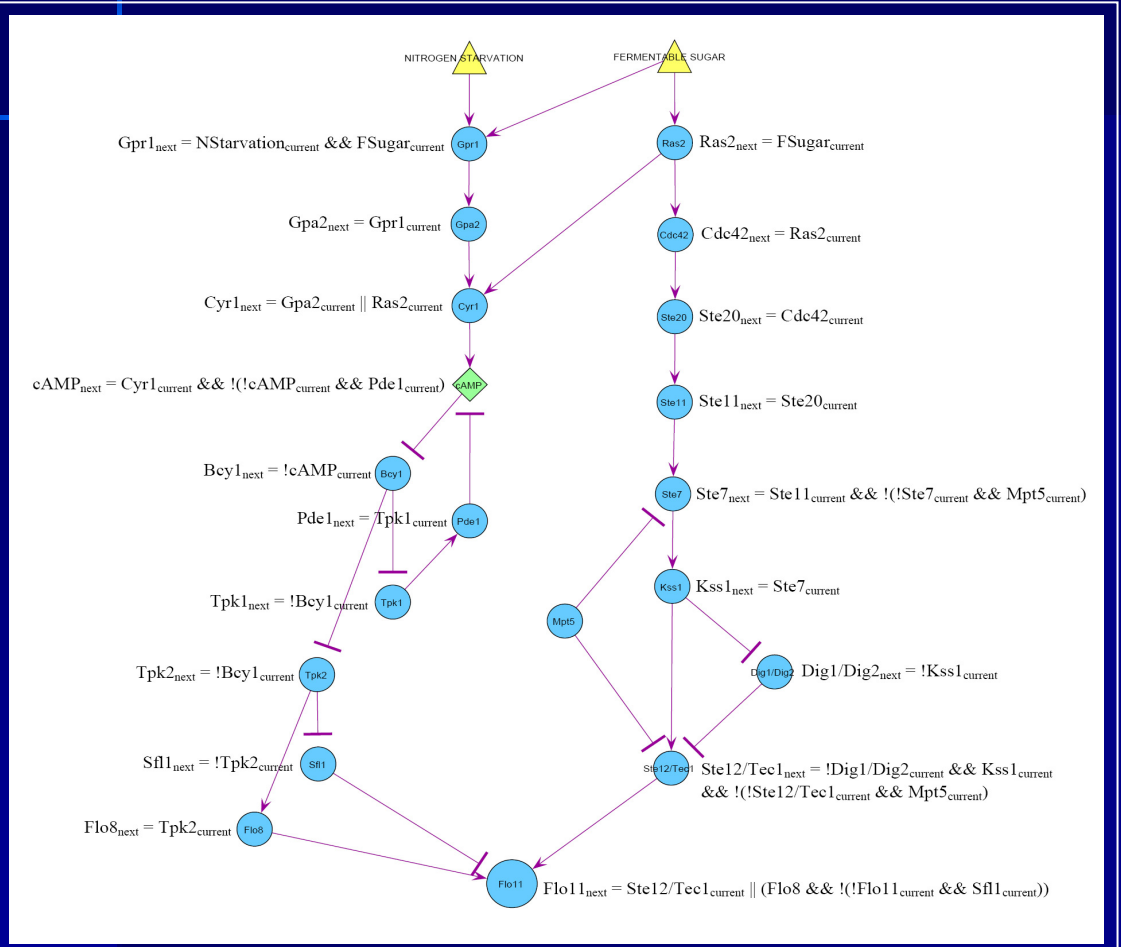
State Space of Boolean Networks

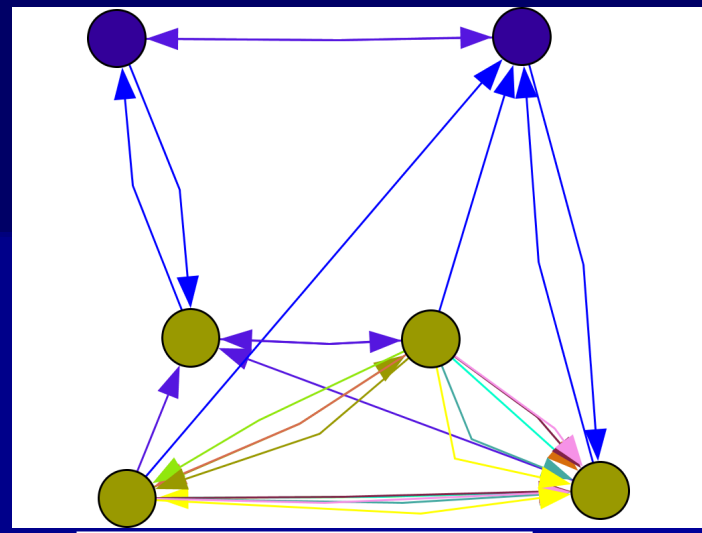


Picture generated using the program DDLab.

- equate cellular states (or fates) with attractors.
- attractor states are stable under small perturbations
  - most perturbations cause the network to flow back to the attractor.
  - some genes are more important and changing their activation can cause the system to transition to a different attractor.

Boolean model of the yeast filamentation network





Non-Filamentous

Filamentous

Environmental Input

Mpt5

Ras2

Cdc42

Ste20

Ste11

Ste7

Kss1

Dig1/2

Tec1-Ste12

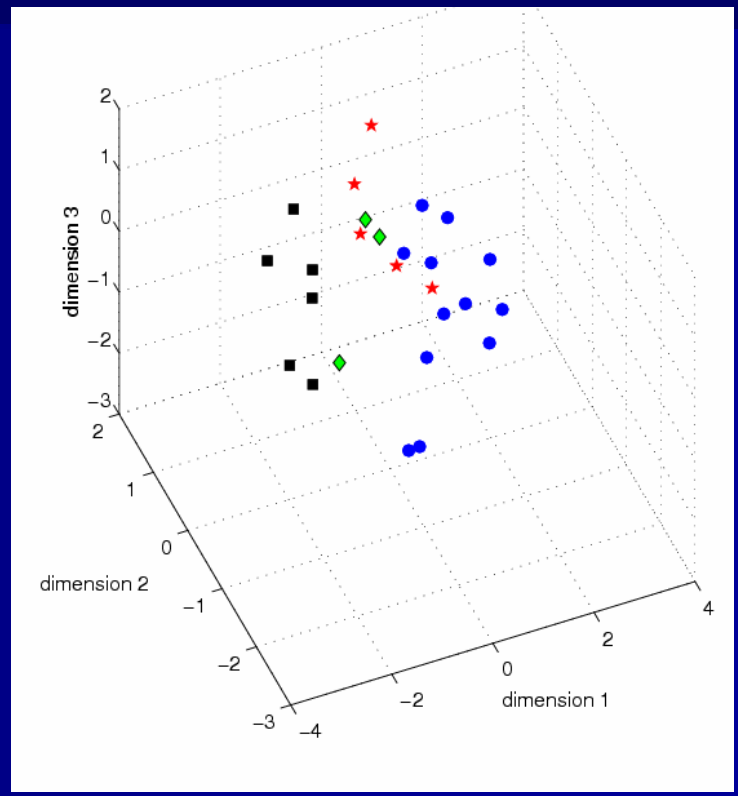
Taylor, Galitski

But can we extract meaningful biological information from gene expression data entirely in the binary domain?

- We reasoned that if genes, when quantized to only two levels (1 or 0) would not be informative in separating known subclasses of tumors, then there would be little hope for Boolean inference of real genetic networks.

Gene expression analysis in the binary domain

- By using binary gene expression data and Hamming distance as a similarity metric, a separation between different subtypes of gliomas is evident, using multidimensional scaling.



Shmulevich, I. and Zhang, W. (2002) *Bioinformatics* 18(4), 555-565.

Boolean Framework

- Limited amounts of data and the noisy nature of the measurements can make useful quantitative inferences problematic and a coarse-scale qualitative modeling approach seems to be justified.
- Boolean idealization enormously simplifies the modeling task.
- We wish to study the collective regulatory behavior without specific quantitative details.
- Boolean networks qualitatively capture typical genetic behavior.

• Albert, R & Othmer, H.G. (2003) *J. Theor. Biol*. **223**, 1-18.
• Mendoza, L., Thieffry, D. & Alvarez-Buylla, R.E. (1999) *Bioinformatics* **15**, 593-606.
• Huang, S. & Ingber, D. E. (2000) *Exp. Cell Res*. **261,** 91-103.
• Li F, Long T, Lu Y, Ouyang Q, Tang C. (2004) *PNAS*. 101(14):4781-6.

BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 261–274

Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks

Ilya Shmulevich[1], Edward R. Dougherty[2], Seungchan Kim[2] and Wei Zhang[1]

[1]Cancer Genomics Laboratory, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 85, Houston, TX 77030, USA and [2]Department of Electrical Engineering, Texas A&M University, College Station, TX 77843, USA

Received on May 2, 2001; revised on July 13, 2001; accepted on October 5, 2001

BIOINFORMATICS

Vol. 18 no. 10 2002
Pages 1319–1331

Gene perturbation and intervention in probabilistic Boolean networks

Ilya Shmulevich[1,*], Edward R. Dougherty[2] and Wei Zhang[1]

[1]Cancer Genomics Laboratory, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 85, Houston, TX 77030, USA and [2]Department of Electrical Engineering, Texas A&M University, College Station, TX 77843, USA

Received on November 13, 2001; revised on March 14, 2002; accepted on March 21, 2002

Probabilistic Boolean Networks (PBN)

- Share the appealing rule-based properties of Boolean networks.
- Robust in the face of uncertainty.
- Dynamic behavior can be studied in the context of Markov Chains.
  - Boolean networks are just special cases.
- Close relationship to (dynamic) Bayesian networks
  - Explicitly represent probabilistic relationships between genes. (Lähdesmäki *et al*. (2006) *Sig. Proc*., 86(4):814-834)
  - Can represent the same joint probability distribution.
- Allow quantification of influence of genes on other genes (stay tuned for examples)

Shmulevich *et al*. (2002) *Proceedings of the IEEE*, 90(11), 1778-1792.

Model Inference from Gene Expression Data

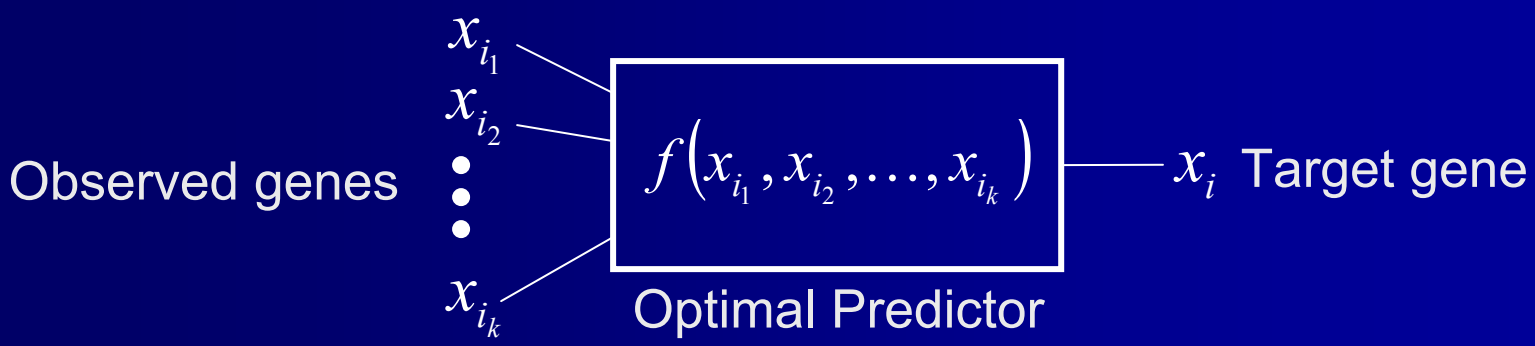
- Two approaches:
  - *Coefficient of Determination* (Dougherty *et al.* 2000)
  - *Best-Fit Extensions*

    Lähdesmäki *et al*. (2003) *Machine Learning*, 52, 147-167.

Coefficient of Determination (COD)

- COD is used to discover associations between variables.
- It measures the degree to which the expression levels of an observed gene set can be used to improve the prediction of the expression of a target gene relative to the best possible prediction in the absence of observations.
- Using the COD, one can find sets of genes related multivariately to a given target gene.

COD Definition



$$\theta = \frac{\varepsilon_i - \varepsilon_{opt}}{\varepsilon_i}$$

ε_i is the error of the best (constant) estimate of x_i in the absence of any conditional variables

ε_{opt} is the optimal error achieved by f

Constraints During Inference

- Constraining the class of predictors can have advantages:
  - lessening the data requirements for reliable estimation;
  - incorporating prior knowledge of the class of functions representing genetic interactions;
  - certain classes of functions are more plausible from the point of view of evolution, noise resilience, network dynamics, etc.

Example of Constraint: Post Classes

Shmulevich *et al.* (2003) *PNAS* 100(19), 10734-10739.

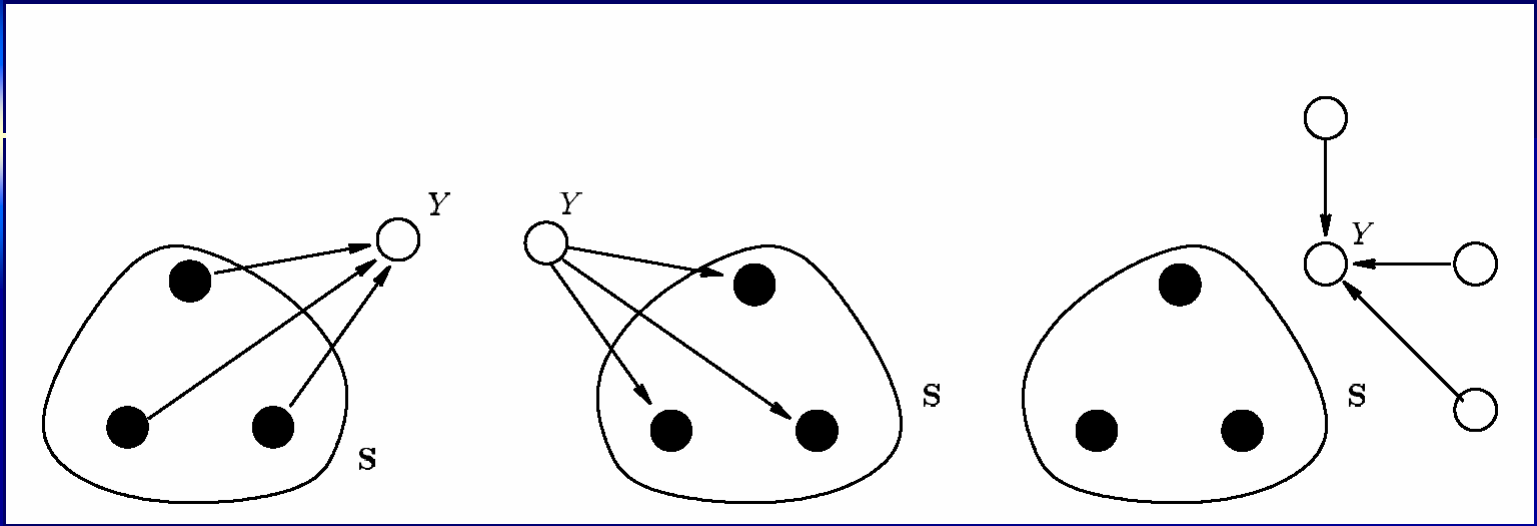
Emil Post (1897-1954)

- The class is sufficiently large (this is important for inference).
- An abundance of functions from this class will tend to prevent chaotic behavior in networks.
- Eukaryotic cells are not chaotic! (Shmulevich *et al.* (2005) *PNAS* 102(38), 13439-13444.)
- Functions from this class have a natural way to ensure robustness against noise and uncertainty.

Subnetworks Theory and Examples

- aim: discover relatively small subnetworks
  - whose genes interact significantly and
  - whose genes are not strongly influenced by genes outside the subnetwork.
- *Principle of Autonomy*
- Start with a 'seed' gene set and iteratively adjoin new genes so as to enhance subnetwork autonomy.

Growing Algorithm



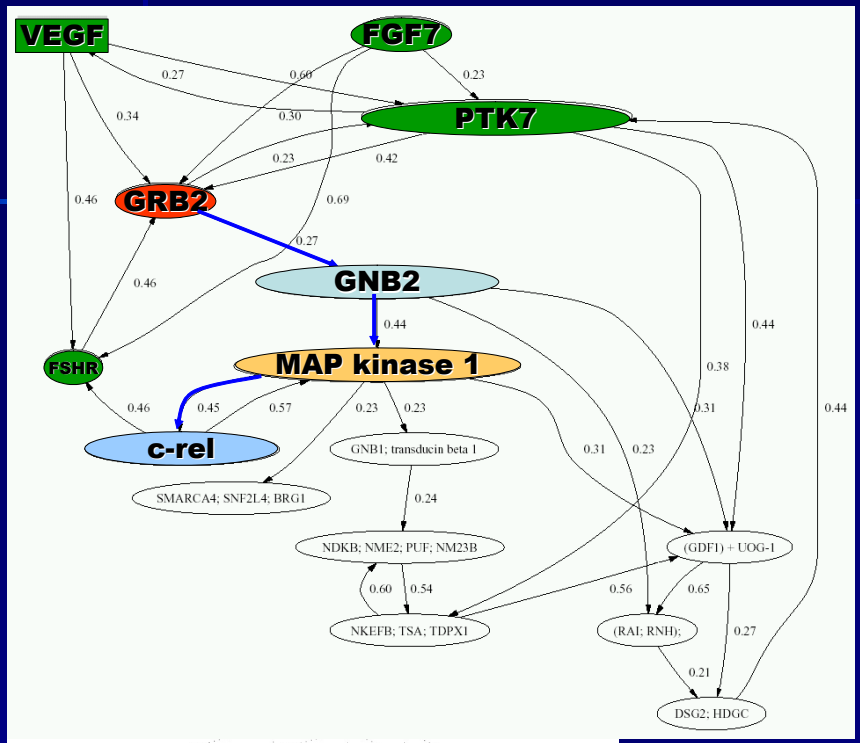
To achieve network autonomy, both of these strengths of connections should be high.

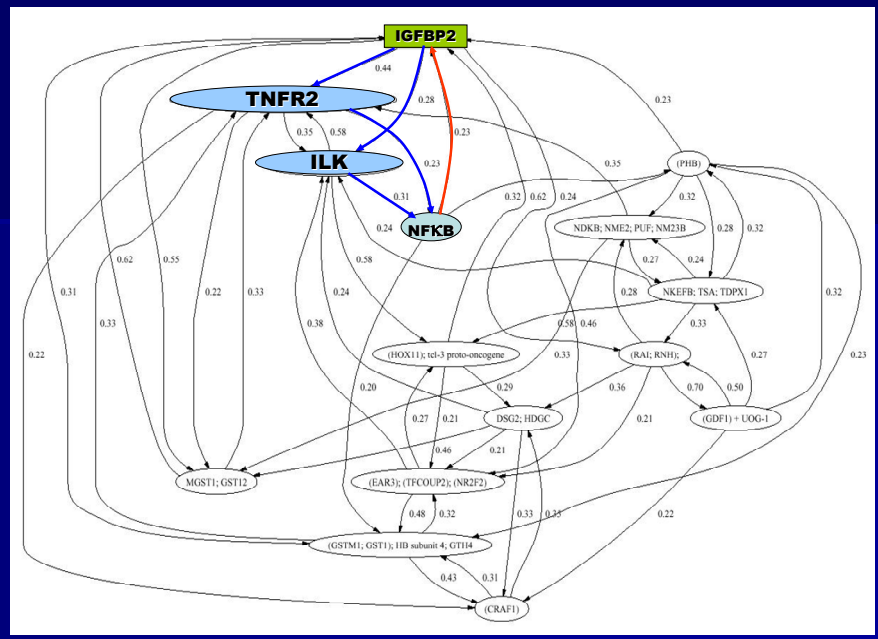
The sensitivity of *Y* from the outside should be small.

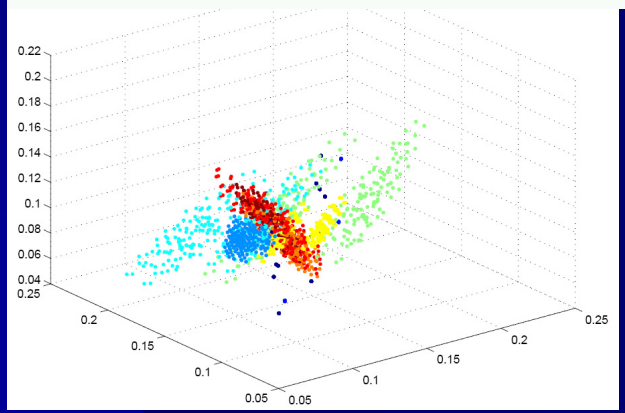
Various stopping criteria can be used

Hashimoto *et al*. (2004) *Bioinformatics* 20(8): 1241-1247.

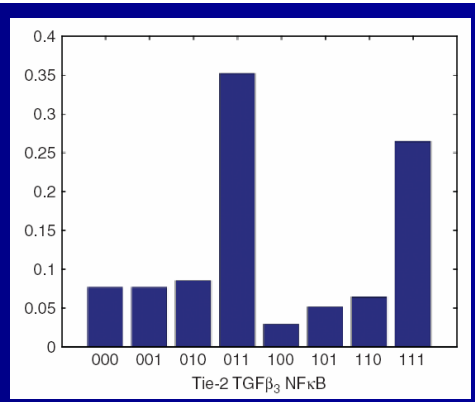
Modeling, Inference, and Simulation of Genetic Networks with Probabilistic Boolean Networks



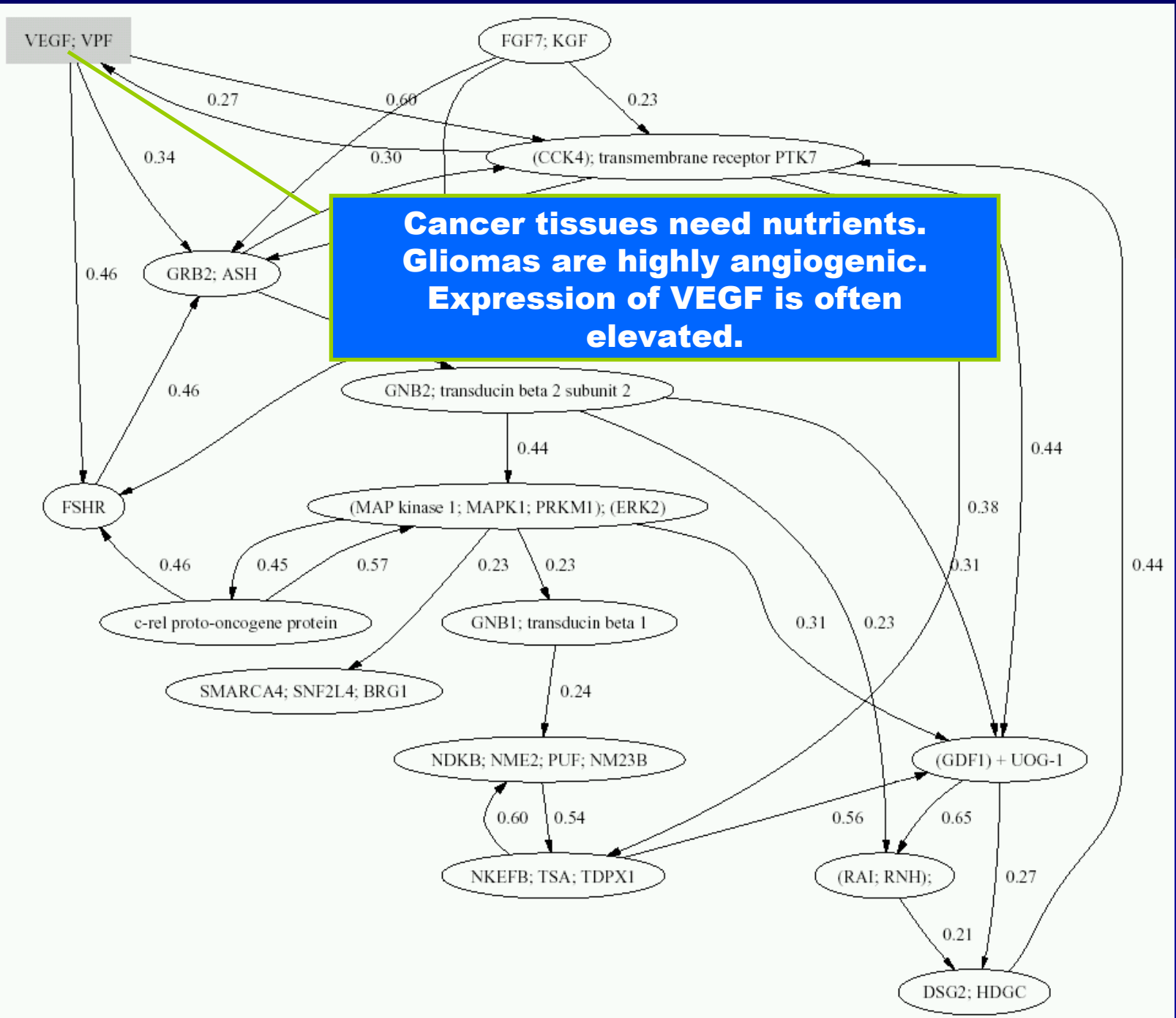




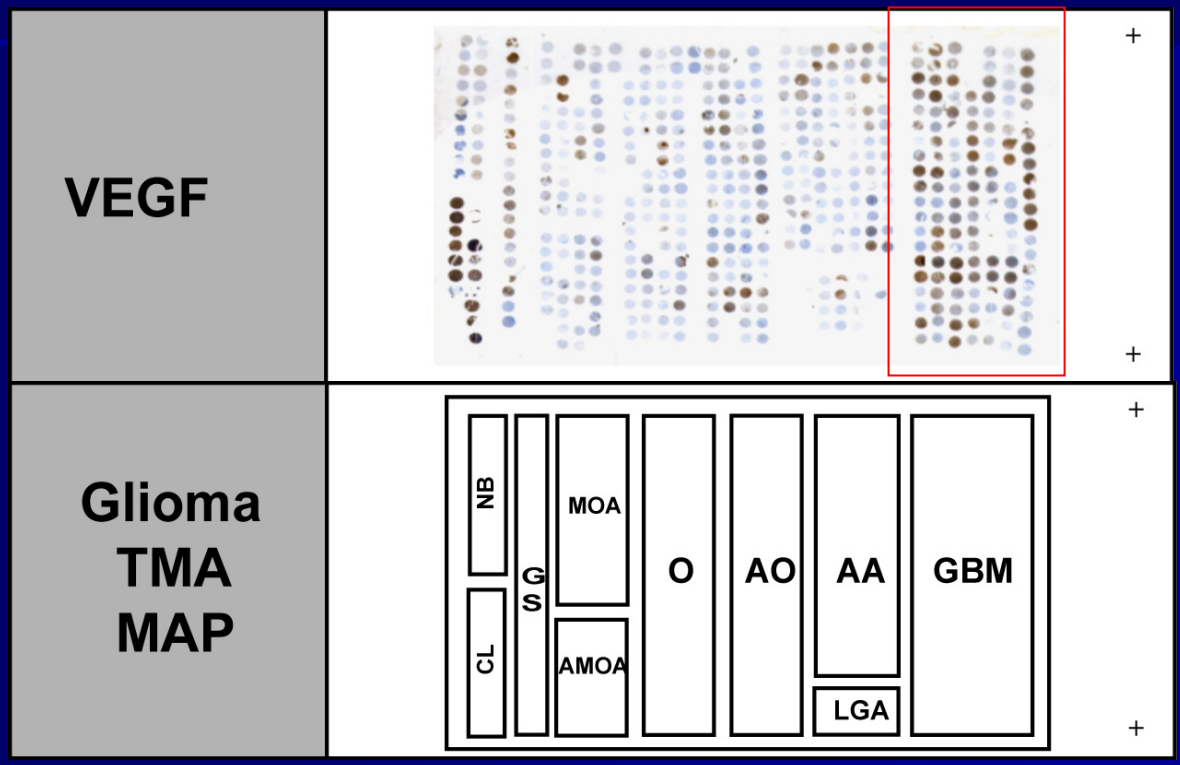
BN/PBN Toolbox

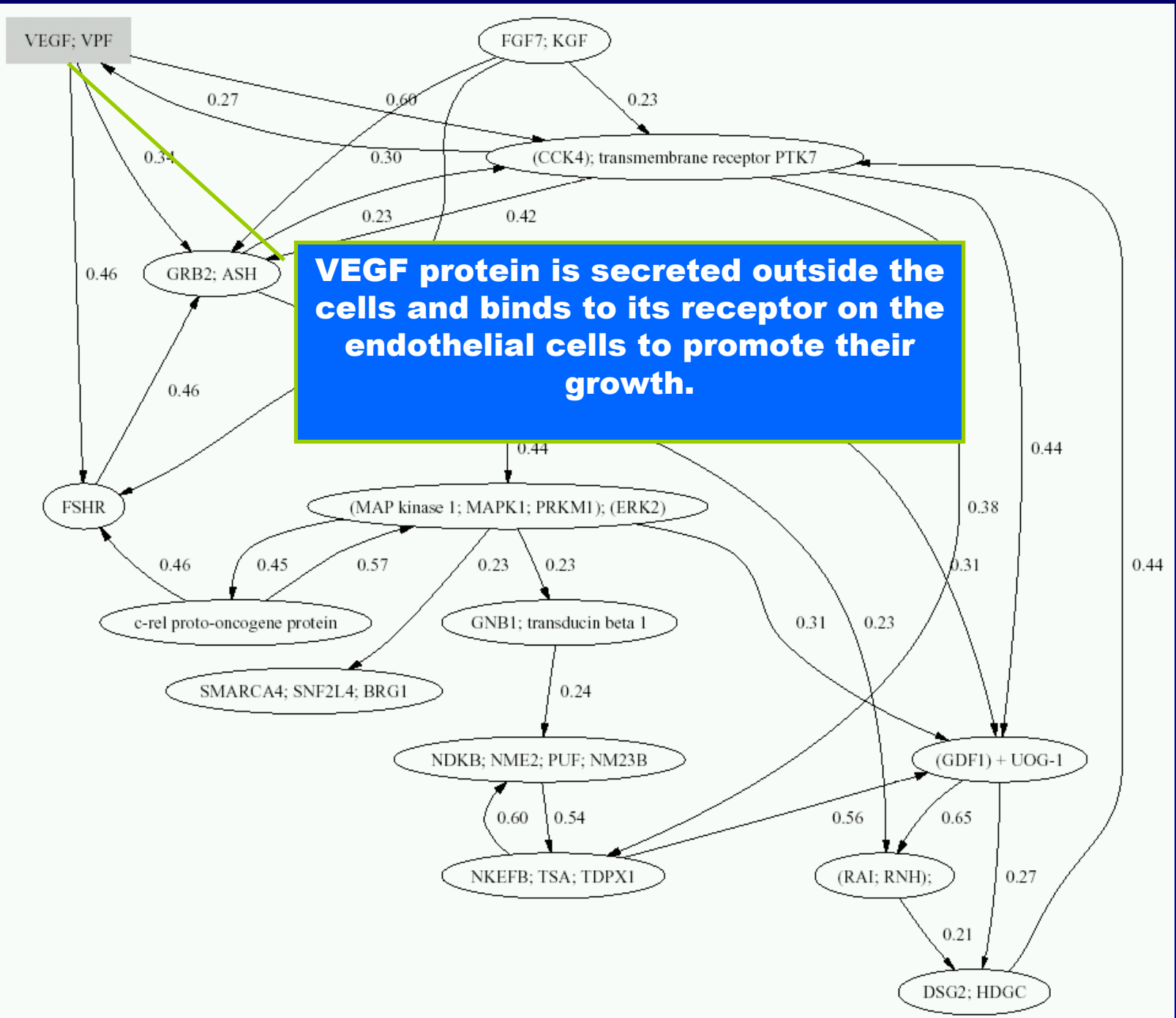


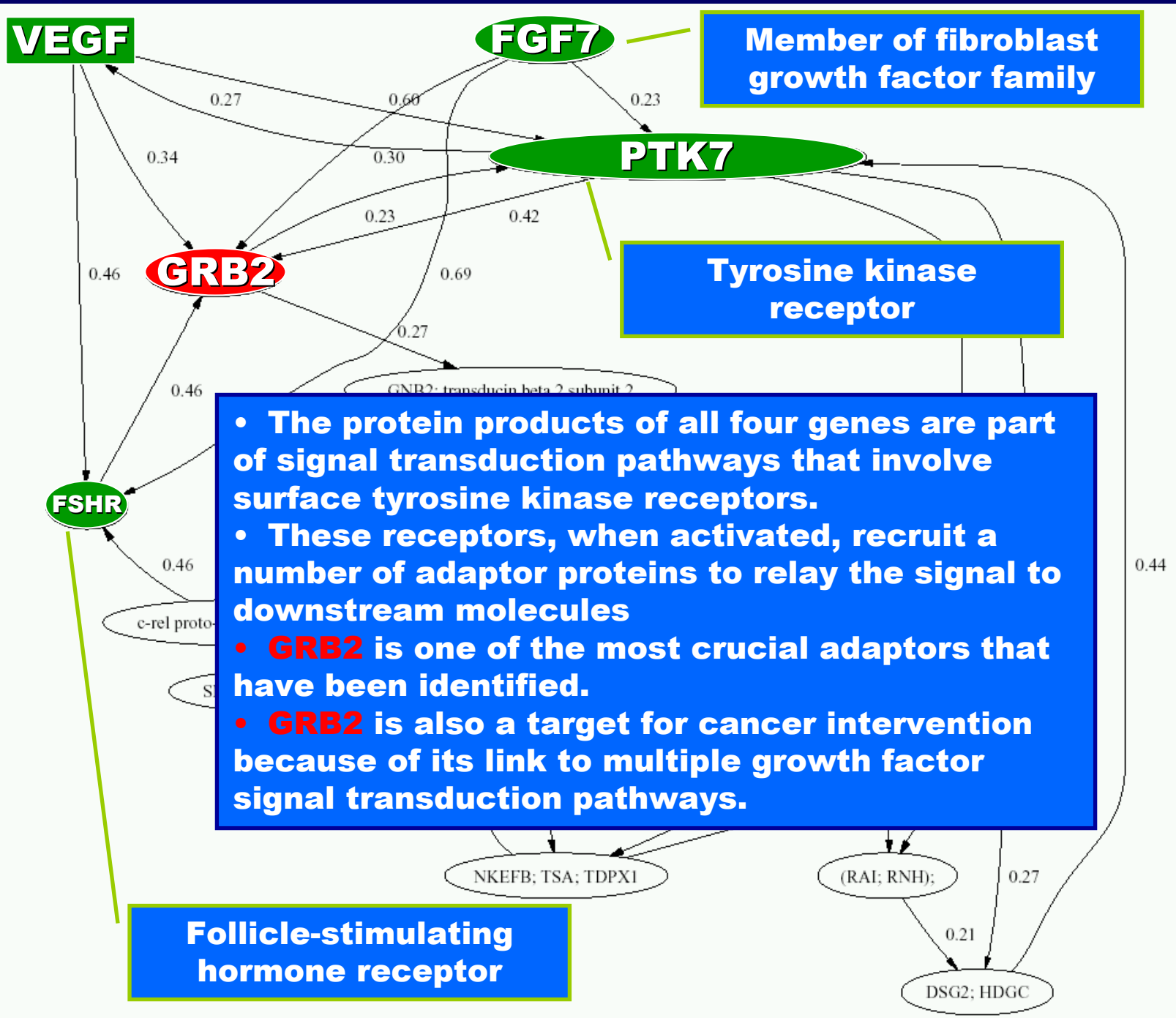
Tie-2	*NFκB*	%	Tie-2	*TGFβ3*	%	*TGFβ3*	*NFκB*	%
OFF	OFF	15.68	OFF	OFF	14.75	OFF	OFF	10.25
OFF	ON	41.58	OFF	ON	42.50	OFF	ON	12.47
ON	OFF	9.21	ON	OFF	7.96	ON	OFF	14.64
ON	ON	31.53	ON	ON	32.78	ON	ON	60.65

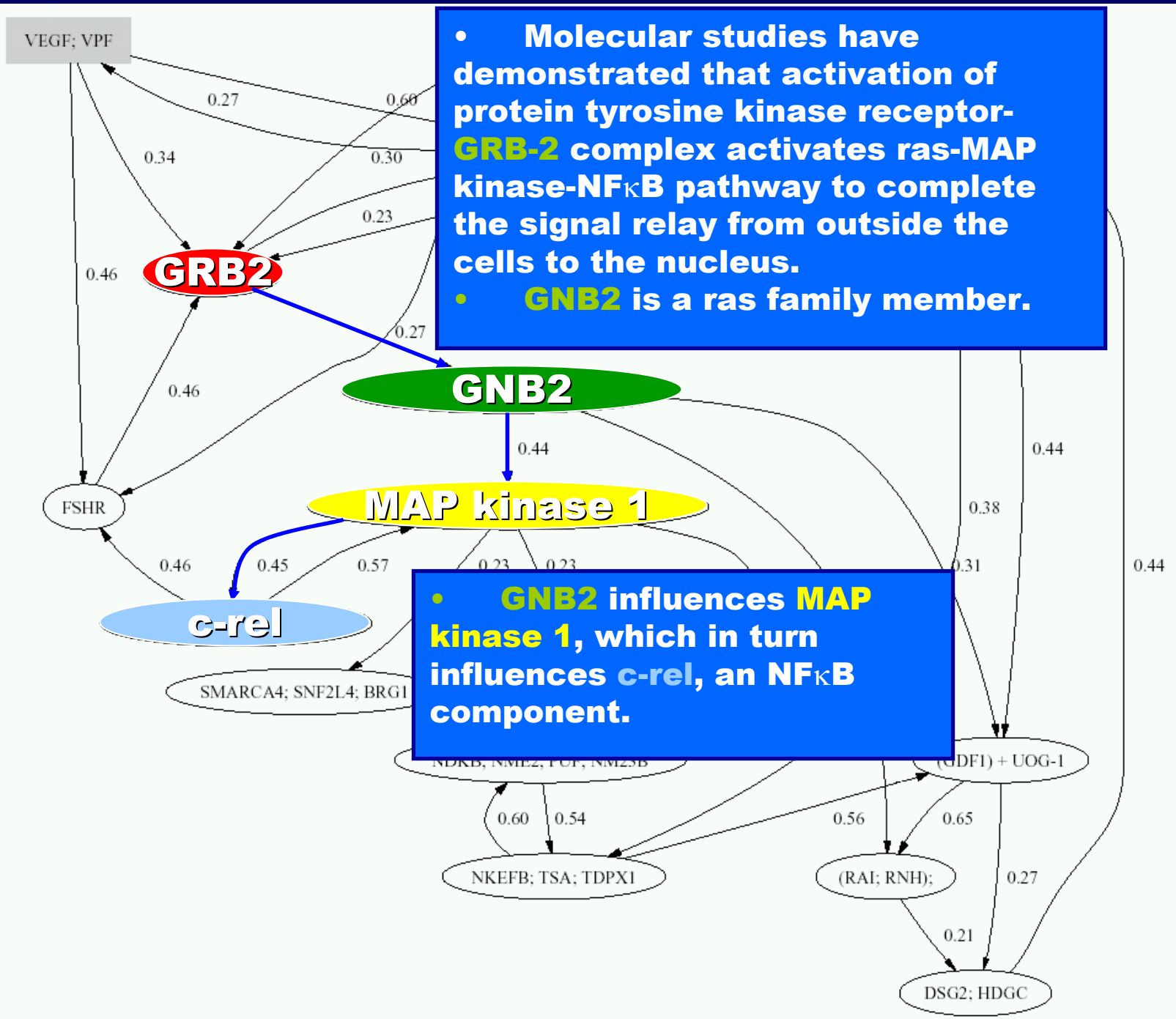

VEGF; VPF
FGF7; KGF
0.27
0.60
0.23
0.34
0.30
(CCK4); transmembrane receptor PTK7
Cancer tissues need nutrients.
Gliomas are highly angiogenic.
Expression of VEGF is often
elevated.
0.46
GRB2; ASH
0.46
GNB2; transducin beta 2 subunit 2
0.44
0.44
FSHR
(MAP kinase 1; MAPK1; PRKM1); (ERK2)
0.38
0.46
0.45
0.57
0.23
0.23
0.31
0.44
c-rel proto-oncogene protein
GNB1; transducin beta 1
0.31
0.23
SMARCA4; SNF2L4; BRG1
0.24
NDKB; NME2; PUF; NM23B
(GDF1) + UOG-1
0.60
0.54
0.56
0.65
NKEFB; TSA; TDPX1
(RAI; RNH);
0.27
0.21
DSG2; HDGC

VEGF is elevated in advanced stage of gliomas
Confirmation and localization by tissue microarray



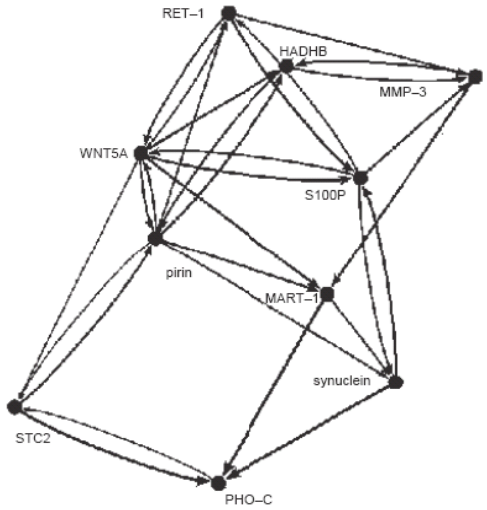

VEGF; VPF
FGF7; KGF
0.27
0.60
0.23
0.34
0.30
(CCK4); transmembrane receptor PTK7
0.23
0.42
0.46
GRB2; ASH
VEGF protein is secreted outside the cells and binds to its receptor on the endothelial cells to promote their growth.
0.46
0.44
0.44
FSHR
(MAP kinase 1; MAPK1; PRKM1); (ERK2)
0.38
0.46
0.45
0.57
0.23
0.23
0.31
0.44
c-rel proto-oncogene protein
GNB1; transducin beta 1
0.31
0.23
SMARCA4; SNF2L4; BRG1
0.24
NDKB; NME2; PUF; NM23B
(GDF1) + UOG-1
0.60
0.54
0.56
0.65
NKEFB; TSA; TDPX1
(RAI; RNH);
0.27
0.21
DSG2; HDGC


VEGF
FGF7
Member of fibroblast growth factor family
0.27
0.60
0.23
0.34
0.30
PTK7
0.23
0.42
Tyrosine kinase receptor
0.46
GRB2
0.69
0.27
0.46
GNB2; transducin beta 2 subunit 2
• The protein products of all four genes are part of signal transduction pathways that involve surface tyrosine kinase receptors.
• These receptors, when activated, recruit a number of adaptor proteins to relay the signal to downstream molecules
• GRB2 is one of the most crucial adaptors that have been identified.
• GRB2 is also a target for cancer intervention because of its link to multiple growth factor signal transduction pathways.
FSHR
0.46
0.44
c-rel proto-
NKEFB; TSA; TDPX1
(RAI; RNH);
0.27
0.21
DSG2; HDGC
Follicle-stimulating hormone receptor



- Molecular studies have demonstrated that activation of protein tyrosine kinase receptor-GRB-2 complex activates ras-MAP kinase-NFκB pathway to complete the signal relay from outside the cells to the nucleus.
- GNB2 is a ras family member.

- GNB2 influences MAP kinase 1, which in turn influences c-rel, an NFκB component.



BIOINFORMATICS

Vol. 18 no. 10 2002
Pages 1319–1331

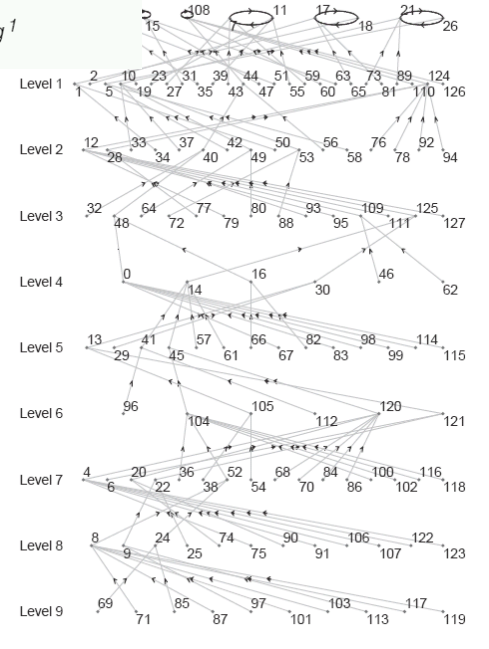
Gene perturbation and intervention in probabilistic Boolean networks

Ilya Shmulevich[1,*], Edward R. Dougherty[2] and Wei Zhang[1]

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 6, JUNE 2006

Optimal Infinite-Horizon Control for Probabilistic Boolean Networks

Ranadip Pal, *Student Member, IEEE*, Aniruddha Datta, *Senior Member, IEEE*, and Edward R. Dougherty, *Member, IEEE*



BIOINFORMATICS

Vol. 20 no. 6 2004, pages 924–930
DOI: 10.1093/bioinformatics/bth008

External control in Markovian genetic regulatory networks: the imperfect information case

Aniruddha Datta[1], Ashish Choudhary[1], Michael L. Bittner[2] and Edward R. Dougherty[1,3,*]

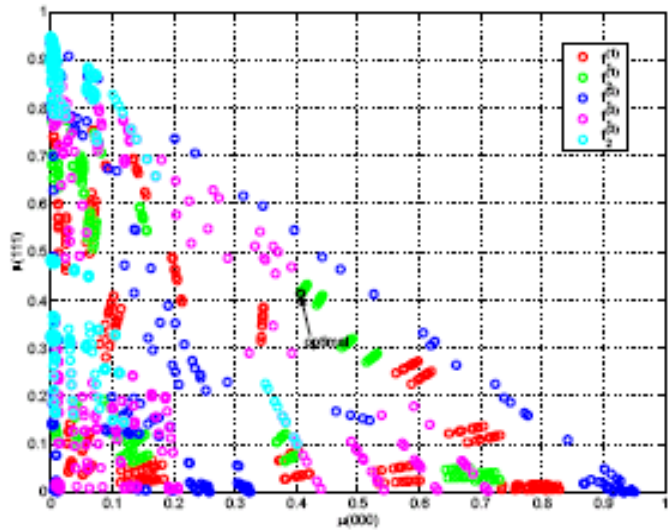
BIOINFORMATICS **ORIGINAL PAPER**

Vol. 21 no. 7 2005, pages 1211–1218
doi:10.1093/bioinformatics/bti131

Systems biology

Intervention in context-sensitive probabilistic Boolean networks

Ranadip Pal[1], Aniruddha Datta[1], Michael L. Bittner[2] and Edward R. Dougherty[1,3,*]



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 22 no. 2 2006, pages 226–232
doi:10.1093/bioinformatics/bti765

Systems biology

Intervention in a family of Boolean networks

Ashish Choudhary[1], Aniruddha Datta[1], Michael L. Bittner[2] and Edward R. Dougherty[1,2,*]

PBN web page

http://personal.systemsbiology.net/ilya/PBN/PBN.htm

- Reprints
- Software (BN/PBN MATLAB Toolbox)
- Posters/Presentations
- Workshops
- Links
- PBN People

PBN Collaborators

Wei Zhang

Harri Lähdesmäki
Olli Yli-Harja
Jaakko Astola

Edward Dougherty
Ronaldo Hashimoto

Marcel Brun
Seungchan Kim
Edward Suh
Huai Li
Michael Bittner

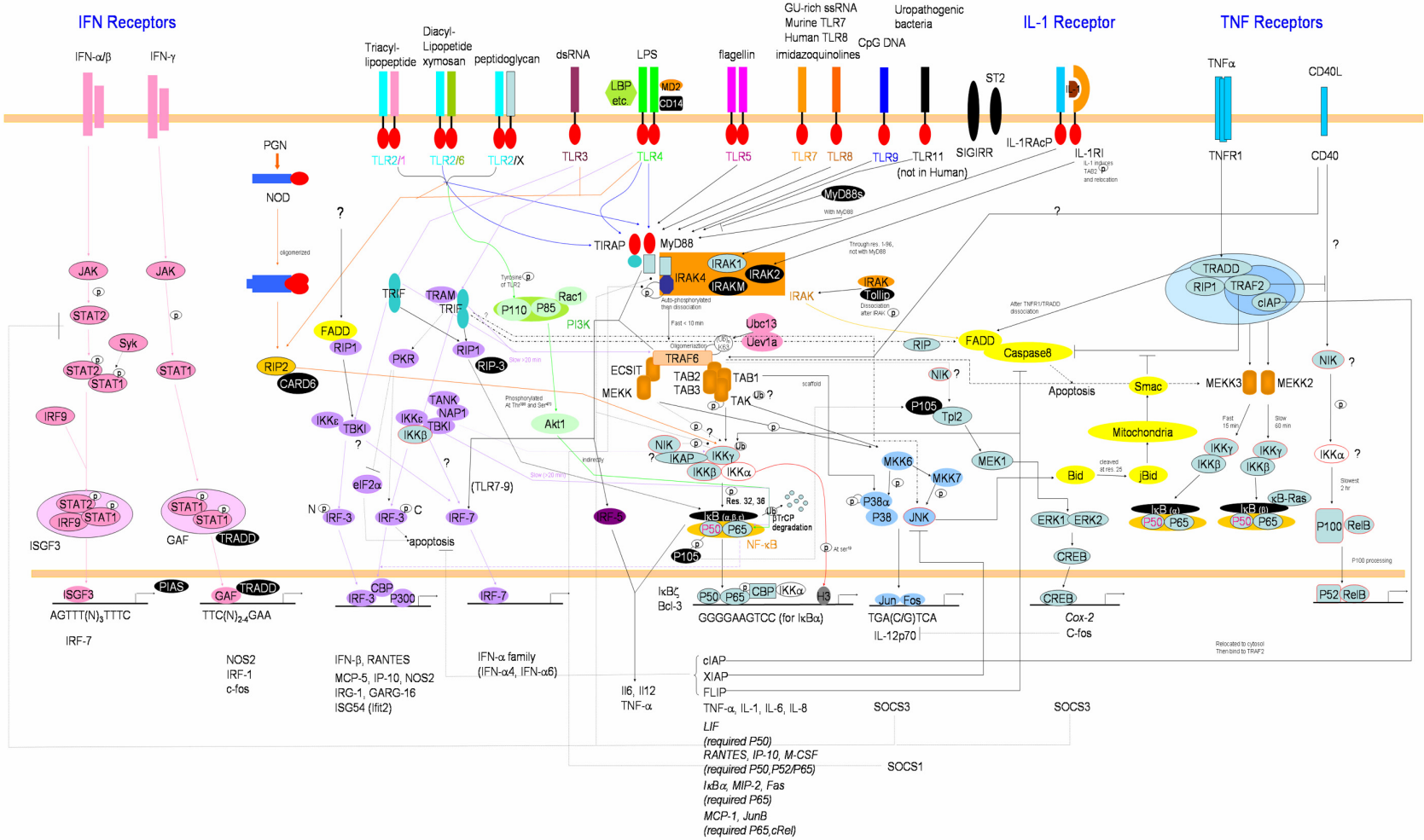
Support

NIH/NIGMS R21 GM070600-01

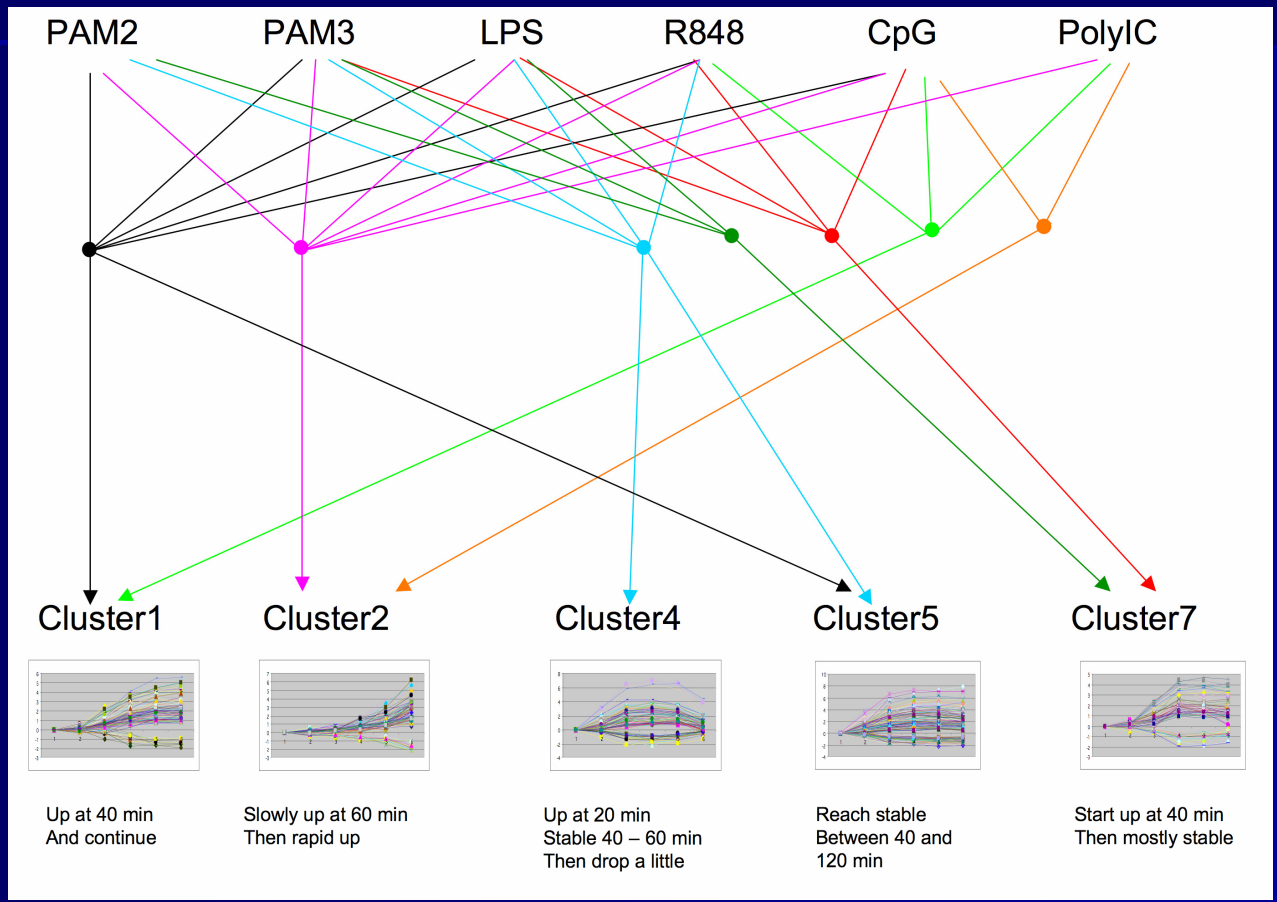
NIH/NIGMS R01 GM072855-01

Integration of Heterogeneous Information

Map of the TLR signaling pathway in the macrophage



Transcriptional responses to TLR ligands are highly coordinated and stimulus dependent



Transcription factor binding site prediction

MotifMogul

[Mogul Homepage] [FAQ] [Example Run] [News] [Comments]

Organism: yeast (S.cerevisiae)

Upload a reference sequence in Fasta Format: GAL10.fa Browse...

Motif Scanners

- [x] Fuzznuc
- [ ] MotifScanner
- [ ] AHAB

Single Scan

- [ ] Verbumculus
- [ ] YMF
- [x] PARS [Palindrome]
- [x] DREAM
- [ ] rmes.poisson
- [ ] rmes.gaussian

Co-regulated

Upload sequences in Fasta Format: GAL10-coregulated.fa Browse...

- [x] AlignACE
- [ ] AlignACE-Long
- [ ] MEME
- [ ] MEME-Long
- [ ] Sampler
- [x] Sampler-Long
- [ ] MotifSampler
- [ ] MotifSampler-Long
- [ ] WConsensus
- [x] WConsensus-Long
- [x] BioProspector
- [ ] BioProspector-Long
- [x] BioProspector-Palindrome

Comparative: Upload a sequence in Fasta Format: GAL10-bayanus.fa Browse...

- [x] Seqcomp
- [x] DBA
- [ ] Bayesaligner

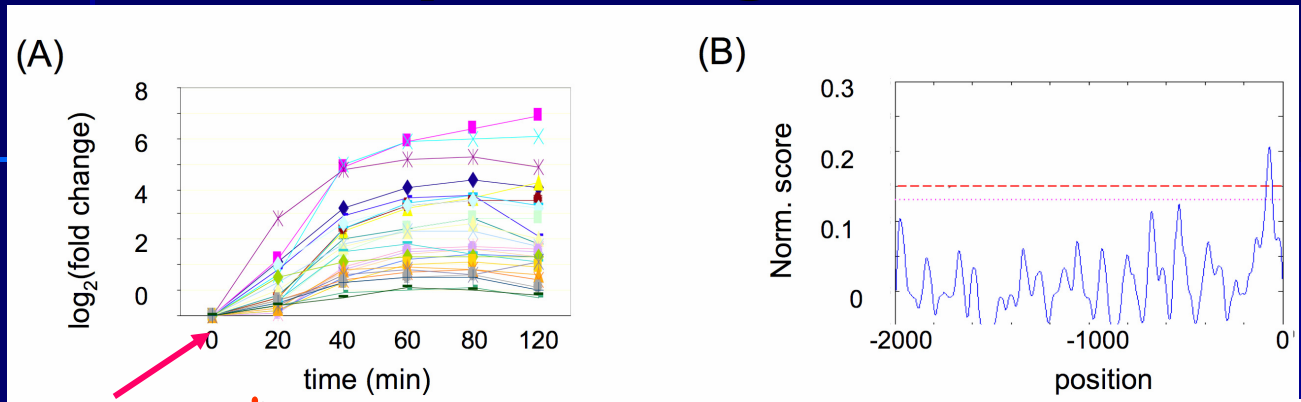
Phylogenetic: Upload sequences in Fasta Format: GAL10-phylo.fa Browse...

- [x] FootPrinter
- [x] MAVID
- [x] AlignACE-Phylo
- [ ] MEME-Phylo
- [ ] Sampler-Phylo
- [ ] MotifSampler-Phylo
- [ ] WConsensus-Phylo
- [ ] BioProspector-Phylo

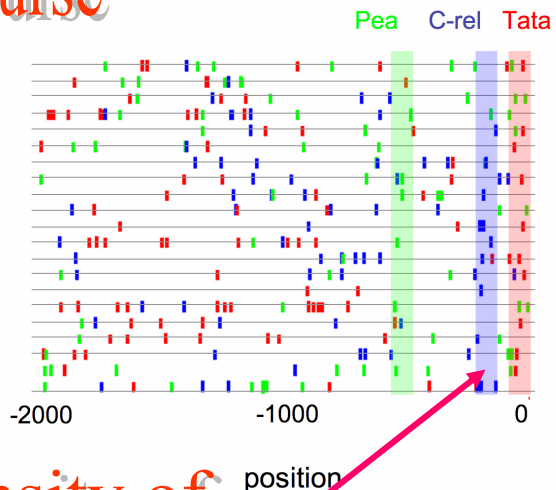
Press - OK - to submit information, or Clear to clear fields.

Alistair Rust, Stephen Ramsey, Mark Robinson

Identifying transcription factors regulating a set of co-expressed genes



Microarray time-course data

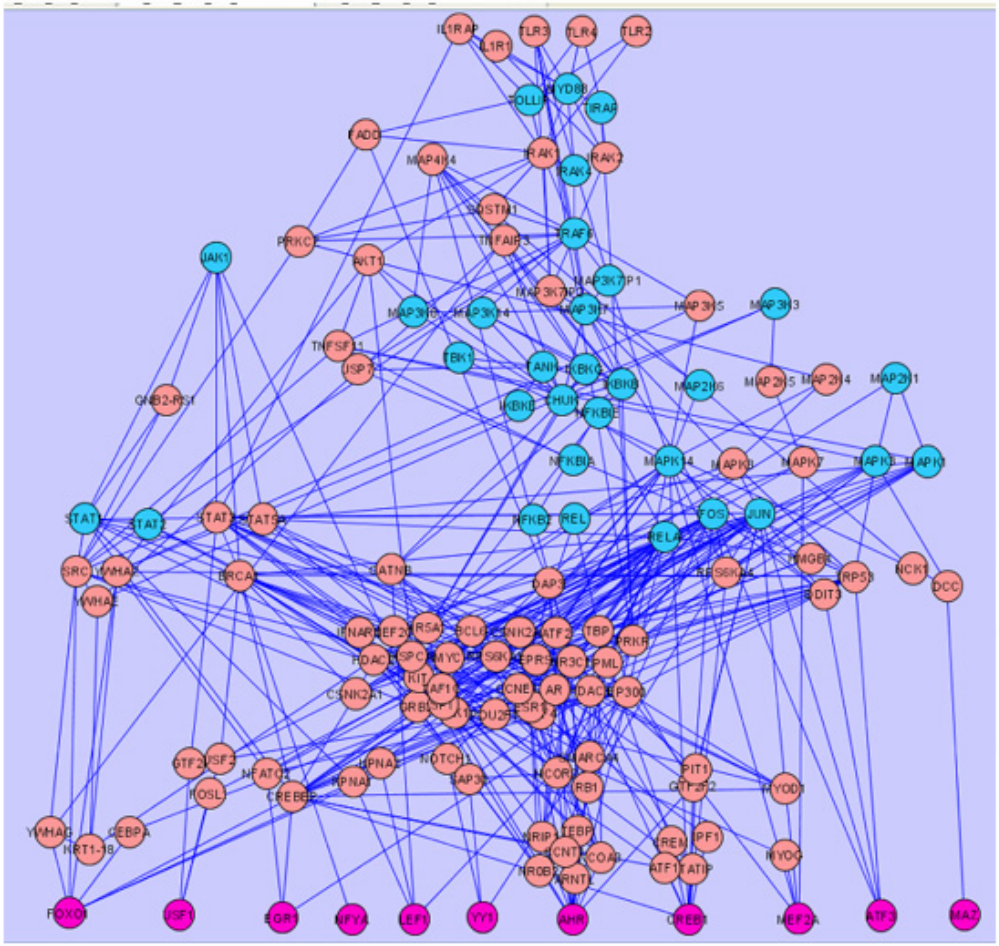


Combined density of predicted TF binding sites

ChIP-chip data

Bin Li

TLR signaling pathway and downstream predicted transcription factors: using protein-protein interactions



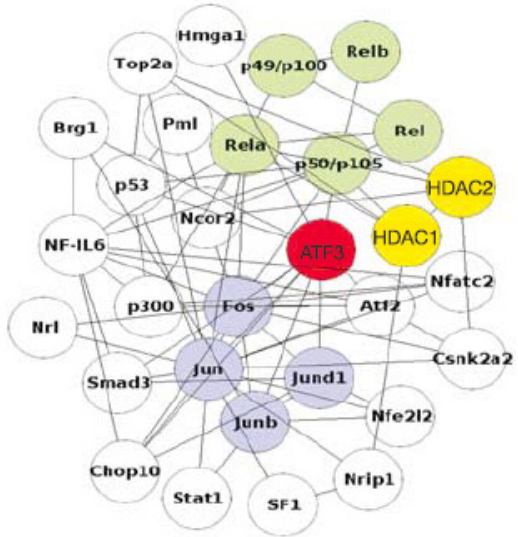
Known TLR Signaling pathway

(Known TFs)

Link to Upstream prot.

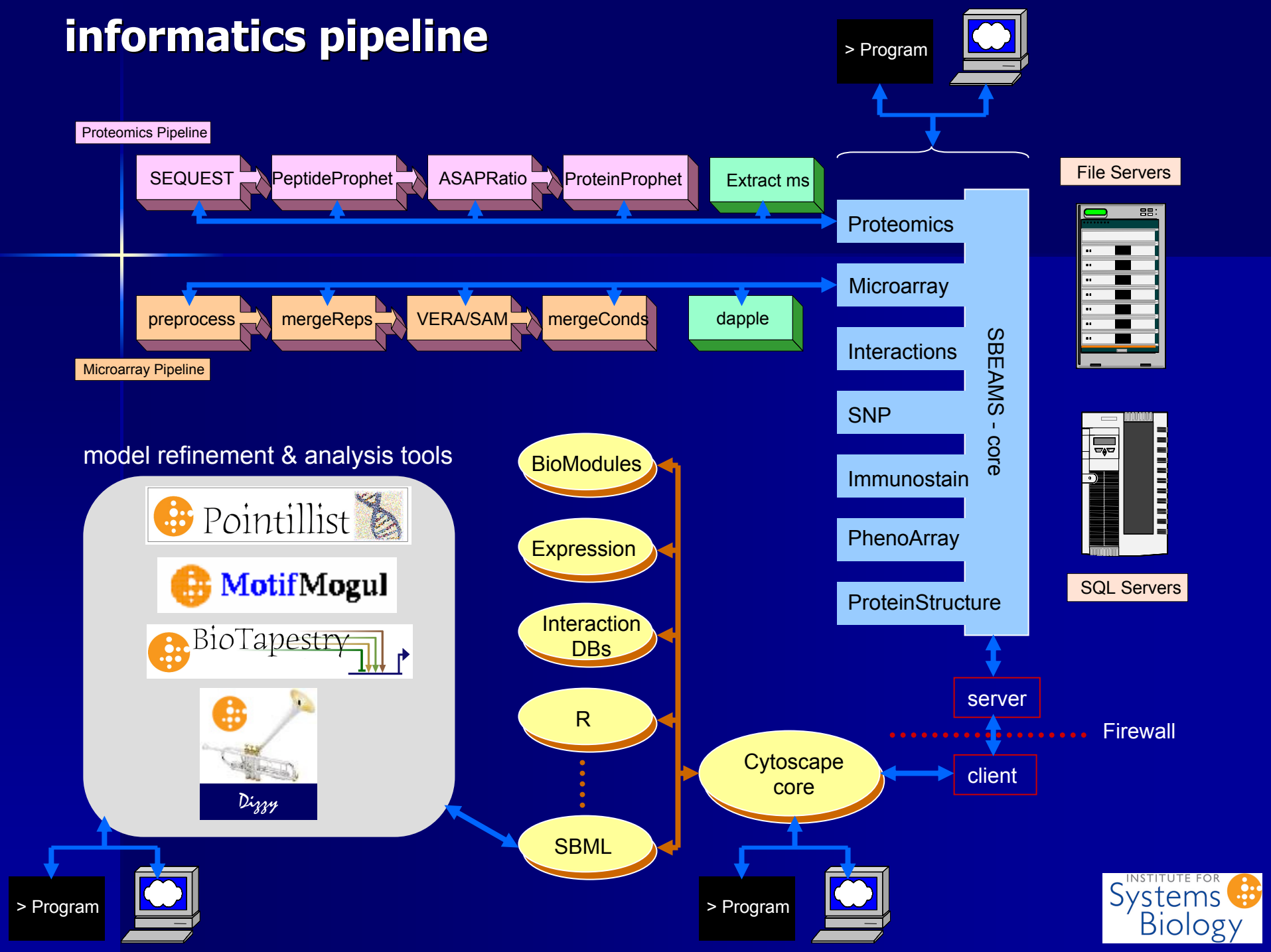
Link to Known TFs

Predicted TFs



ATF3 (red) is predicted to interact with a number of TFs, including members of the AP1 (light blue) and NF-kB (light green) TF complexes.

informatics pipeline



SBEAMS - Systems Biology Experiment Analysis Management System

Primary

Login: **edeutsch** (2) Group: Admin

Project: agingras - TAP_tag_pulldown_HEK293

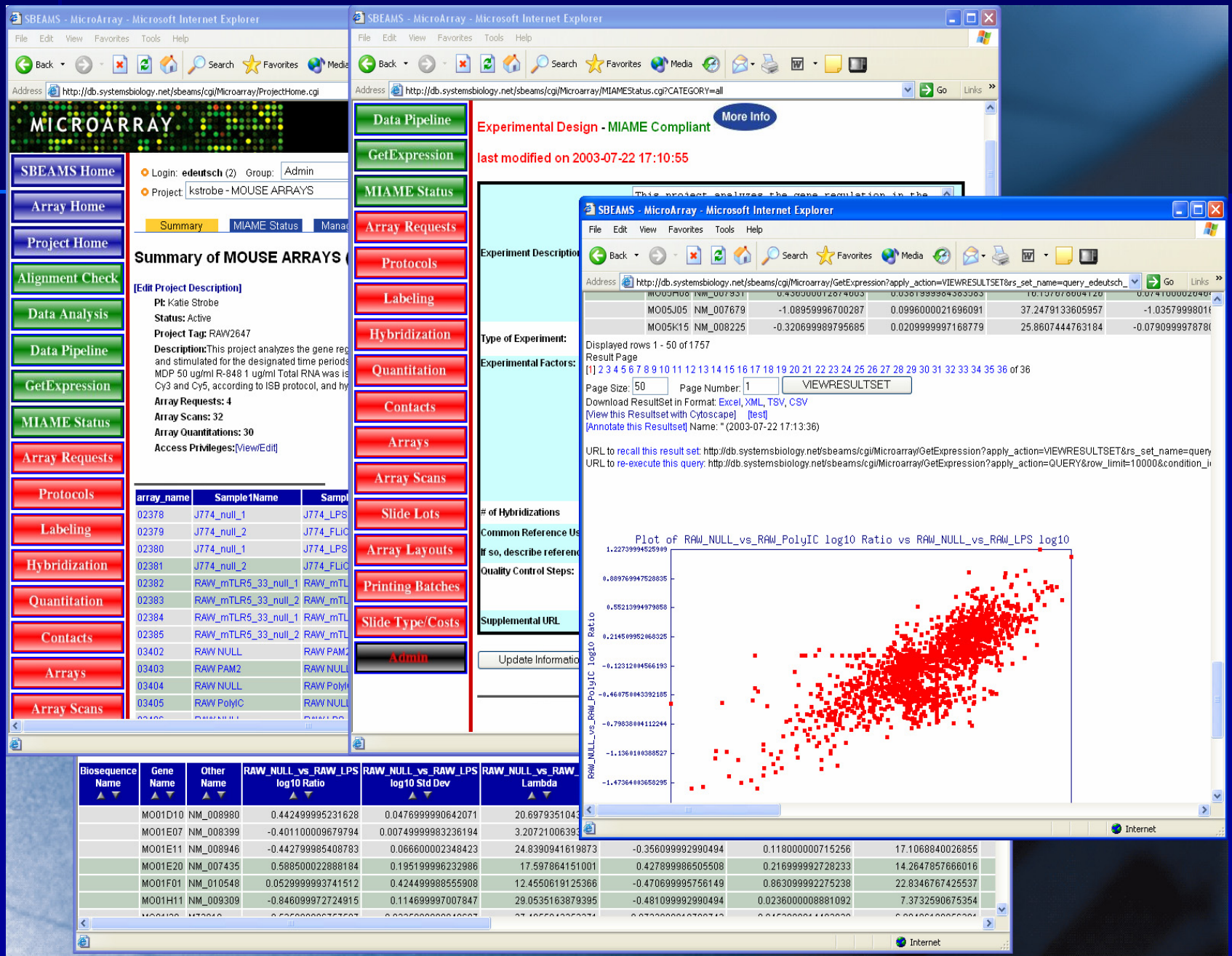
Welcome to the Systems Biology Experiment Analysis Management System (SBEAMS) interface. Note your current work group and project listed above and change it via the pull down menus if desired.

The modules available to you are listed below and on the navigation bar on the left. Select one of the available modules by clicking clicking on it to see the home page for that module.

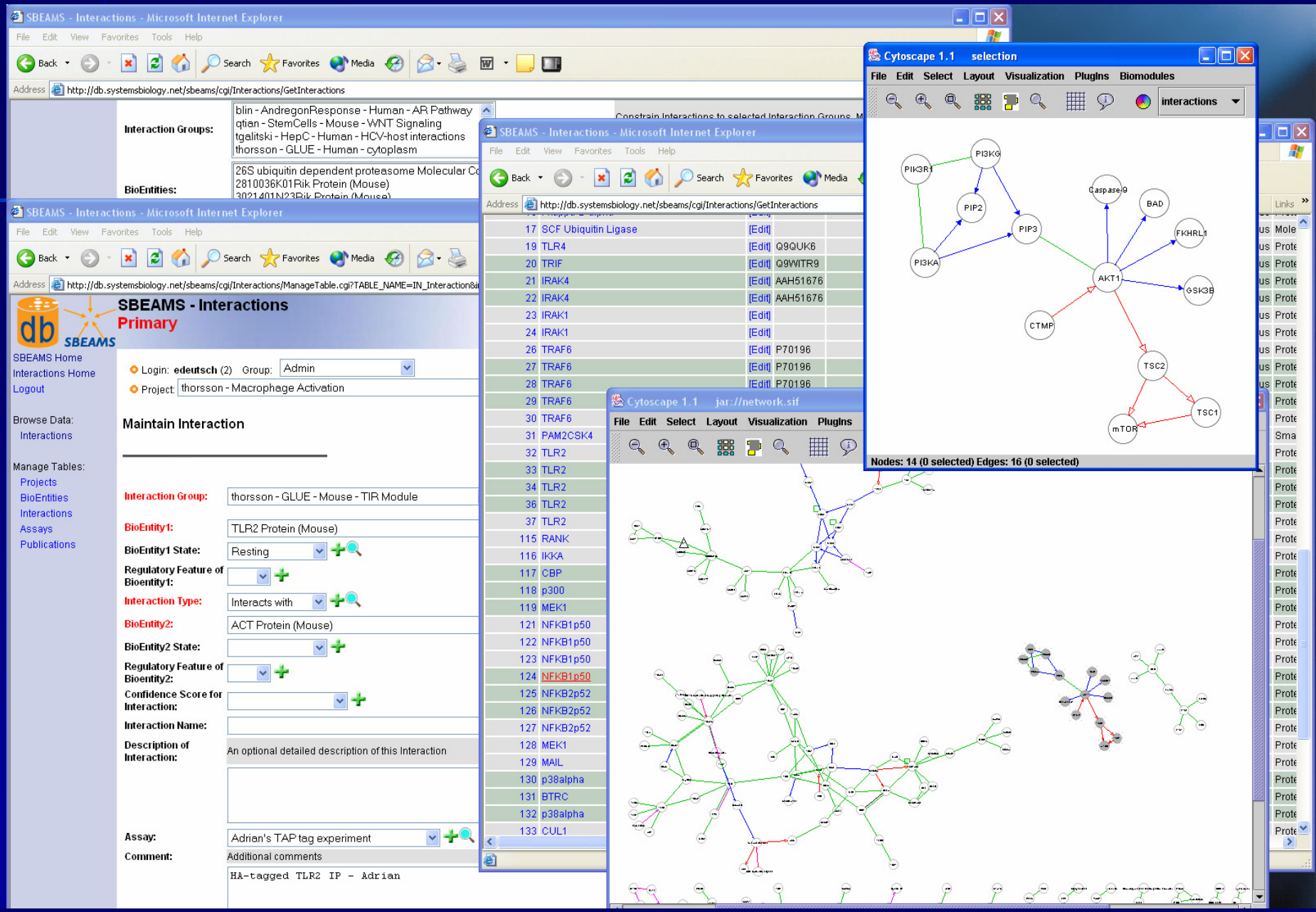
This system is still under active development. Please report bugs, problems, difficulties, and suggestions to **edeutsch@systemsbiology.org**.

- SBEAMS - Microarray
- SBEAMS - Proteomics
- SBEAMS - BEDB
- SBEAMS - BioLink
- SBEAMS - Biosap
- SBEAMS - Cytometry
- SBEAMS - GEAP
- SBEAMS - Immunostain
- SBEAMS - Inkjet
- SBEAMS - Interactions
- SBEAMS - PhenoArray
- SBEAMS - ProteinStructure
- SBEAMS - SNP
- SBEAMS - UESC
- SBEAMS - Tools

SBEAMS – Microarray



SBEAMS – Interactions



Paul Shannon, Eric Deutsch, Martin Korb, Vesteinn Thorsson, Kelly Smith

Directions (Cytoscape)

- Model-based perturbation analysis
  - how does transient or permanent perturbation affect the other nodes and what is the long-run behavior of the network?
- Intervention and control
  - what is the optimal intervention strategy for a particular steady-state behavior?
- Integration of inference engines
  - BNs, PBNs, DBNs, etc.
- Steady-state analysis
  - what is the joint steady-state distribution of a group of genes?

Directions (Cytoscape)

- Viewing state-space (dynamics)
  - deterministic or probabilistic
- Viewing transitions between steady-states (attractors)
  - which gene perturbations cause a transition from one attractor to another?
- Functionality within an enterprise (distributed component-based) architecture (web services)

acknowledgments

- Bin Li
- Daehee Hwang
- Bill Longabaugh
- Martin Korb
- Stephen Ramsey
- Natalya Yudkovsky
- Alistair G. Rust
- Vesteinn Thorsson
- Mark Gilchrist
- Kathleen Kennedy
- Jared Roach
- Hamid Bolouri
- Alan Aderem

Support:
U54 AI54253 NIH/NIAID

P50 GM076547 NIH/NIGMS