Modelling of Biological Systems

Satya Swarup Samal

B-IT, University of Bonn

Types of Modelling Approaches

- Data driven modelling
  - Model the responses (labels) i.e. biological states or clinical outcomes with respect to the observables e.g. molecular or clinical data e.g. supervised methods.
  - Responses or labels not given e.g. unsupervised methods.
- Mechanistic modelling
  - Biological systems as a *complex non-linear dynamical network* of molecular parts.
  - Facilitate quantitative predictions of emergent behaviour of networks.
  - Study casual relationships. Hence, control and manipulation of biological processes possible.

Data Driven Modelling: Motivation

- Metabolic network as enzyme catalysed biochemical reactions (e.g. KEGG).
- Such networks are used to analyze and understand human diseases based on -omics data (e.g. enrichment methods applied to gene sets).
  - The underlying reaction network structure is usually ignored.
- Our approach computes the sub-reaction systems (pathways) and computes its statistical association with –omics data derived from different clinical phenotypes .

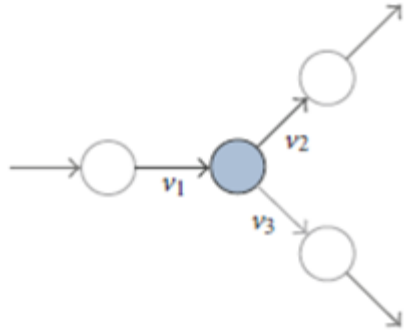
Predictors/Features

$$Y = f(X) + \epsilon.$$

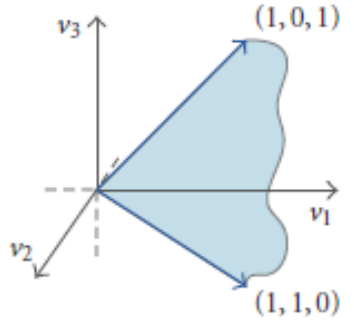
Responses Errors

Pathway Enumeration

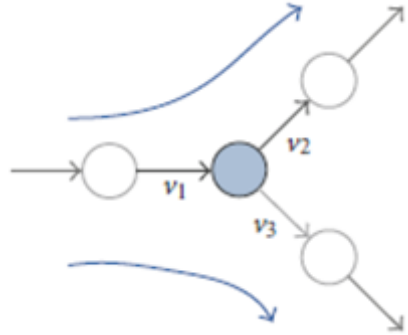
- Decompose network into pathways in an unbiased manner using algebraic techniques.
  - Depends on the structure of the network and is invariant.
  - In literature, such pathways referred as **Extreme Currents (ECs)**, Extreme Pathways, Elementary Flux Modes.
- May grow exponentially with size of network.
  - Infeasible for very large networks e.g. genome scale models.
- Such pathways have many applications e.g. drug target identification, network robustness analysis, etc (Papin et al. 2003).



1. Steady State of chemical species (Equations)
2. Non-Negativity of reaction fluxes (Inequalities)



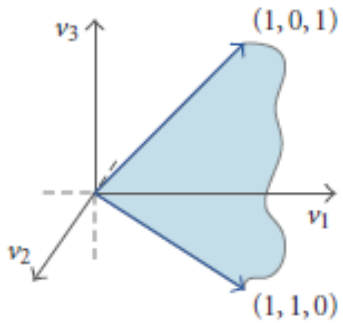
Flux Cone defined by set of inequalities



ECs in blue thick arrows

Image adopted from (Llaneras and Picó, 2010)

Network Features



Reaction Fluxes

ECs

1	0	3

1. Enumerate the ECs.

Genes

ECs

1	0	0

2. Gene Sets: Map non-zero entries in EC vector to genes

Addresses correlation due to network structure

Network Feature 1

EC_1 , EC_2

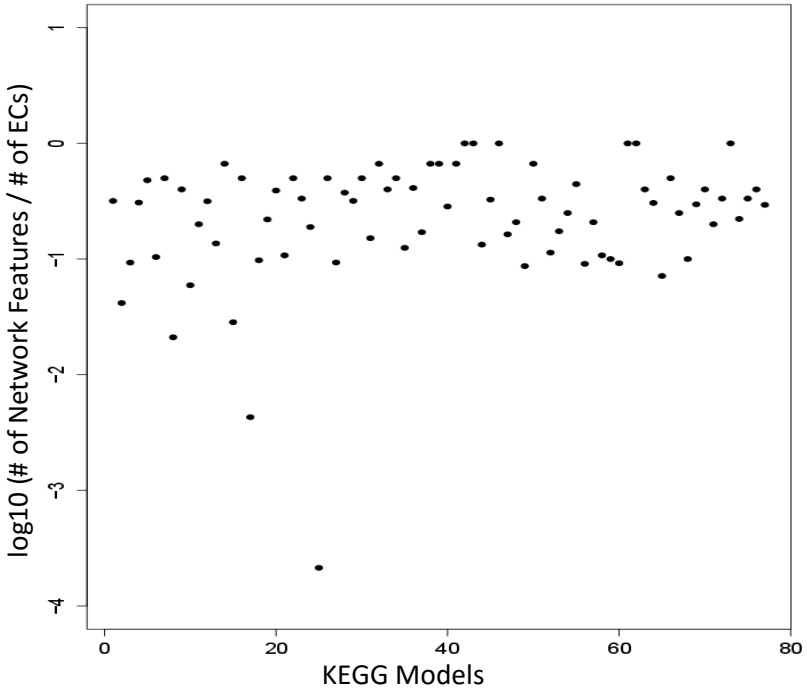
Network Feature 2

EC_5 , EC_6

Network Feature 3

EC_3 , EC_4

3. Network Features: Average Linkage Clustering of gene sets (Jaccard index as similarity measure). The union of elements in a cluster of ECs



Network Features with Gene Expression Data

- Summarise expression of a gene set corresponding to a network feature into an activity score.
  - Take first principal component from Principal Component Analysis.
  - Takes into account variance of the data.
  - Similar approach in Bild et al. 2006.
- Generate feature matrix

Network Feature 1

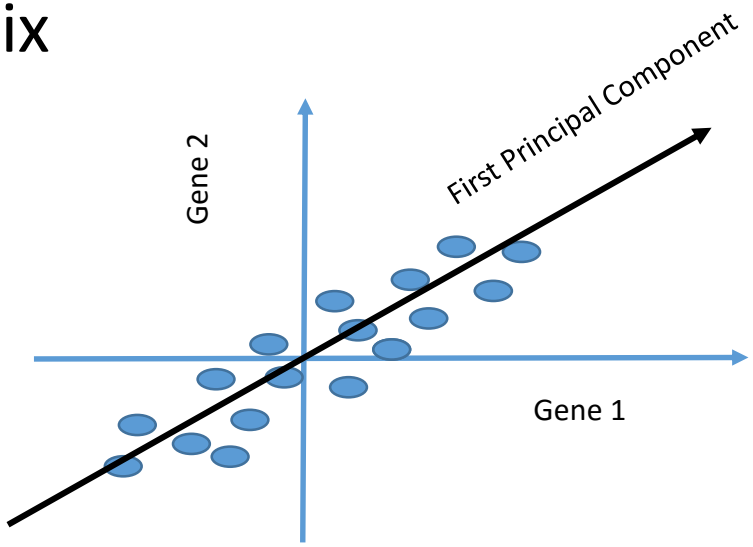
EC_1 , EC_2

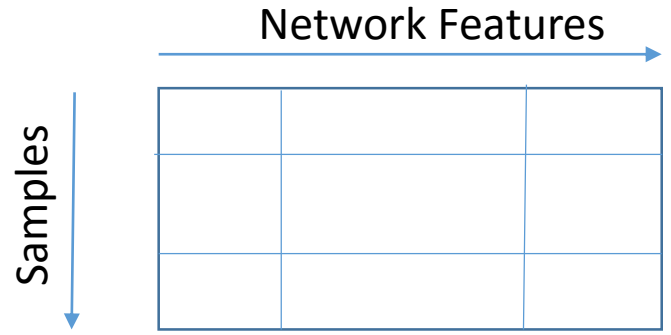
Network Feature 2

EC_5 , EC_6

Network Feature 3

EC_3 , EC_4





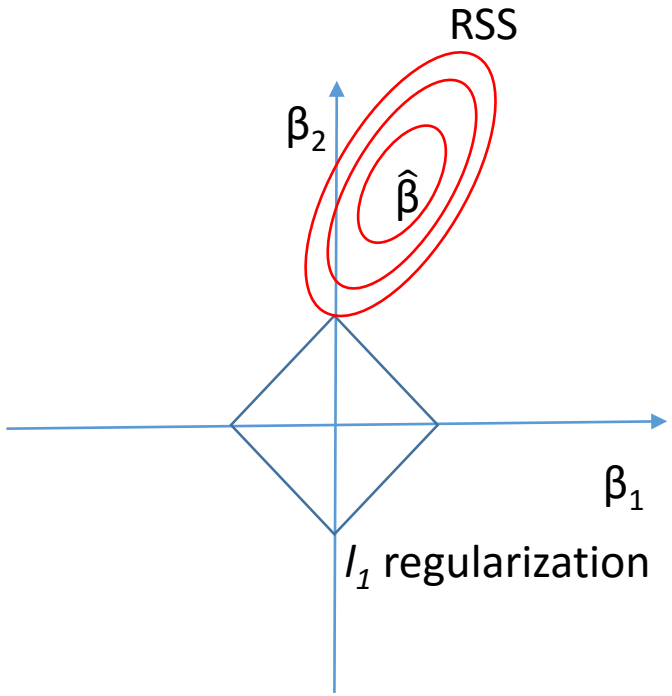
Sparse Group Lasso (SGL)

- Selection of Phenotype Associated Network Features via SGL
- Linear model with regularization i.e. shrinks the coefficient to zero (Simon et al. 2013).
- Optimization Function:

$$\min_\beta \frac{1}{2n} \left|\left| y - \sum_{l=1}^{m} X^{(l)} \beta^{(l)} \right|\right|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l} ||\beta^{(l)}||_2 + \alpha\lambda ||\beta||_1$$

Residual Sum of Errors (RSS) | Group-wise Sparsity | Overall Sparsity

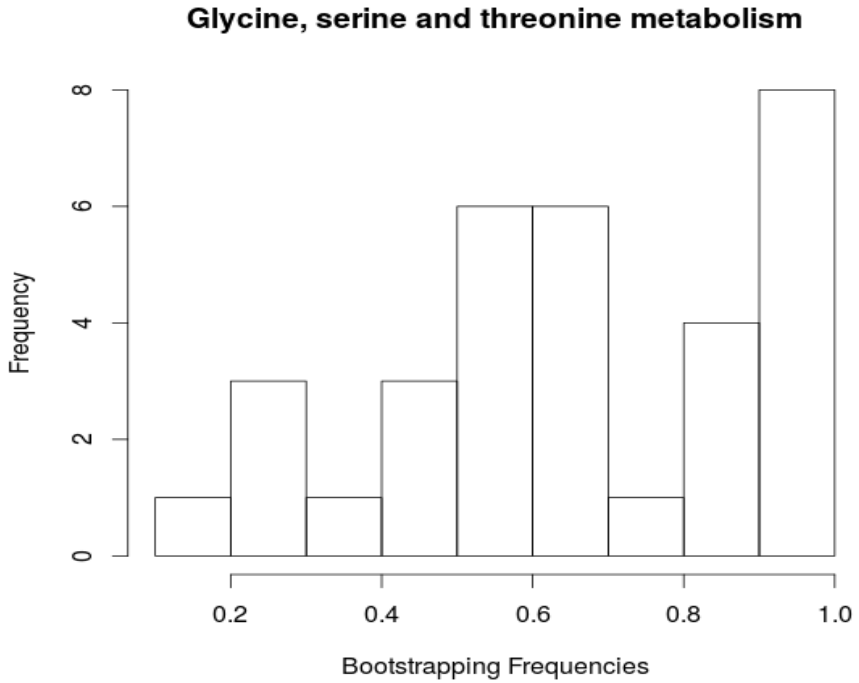
- λ the tuning parameter controlling the sparsity of coefficient vector.
- α is a parameter that balances between sparse selection of whole feature groups and sparse selection within each feature group.
- Features are assigned to *m* groups (defined by average linkage clustering) based on correlation in gene expression data (Bühlmann et al. 2013).



Addresses correlation in data

Case Study I: Prostate Cancer

- A comprehensive analysis of pathways in prostate cancer was reported in (Sreekumar et al., 2009).
  - Sarcosine was found to be highly elevated in tumor samples.
- Pathway: Sarcosine mappable to glycine, serine and threonine pathway in KEGG database.
- -omics data: gene expression data from Brase et al., 2011, comprising 47 prostate tumor tissue samples and 48 normal prostate tissue samples.
- Overview of the pathway:

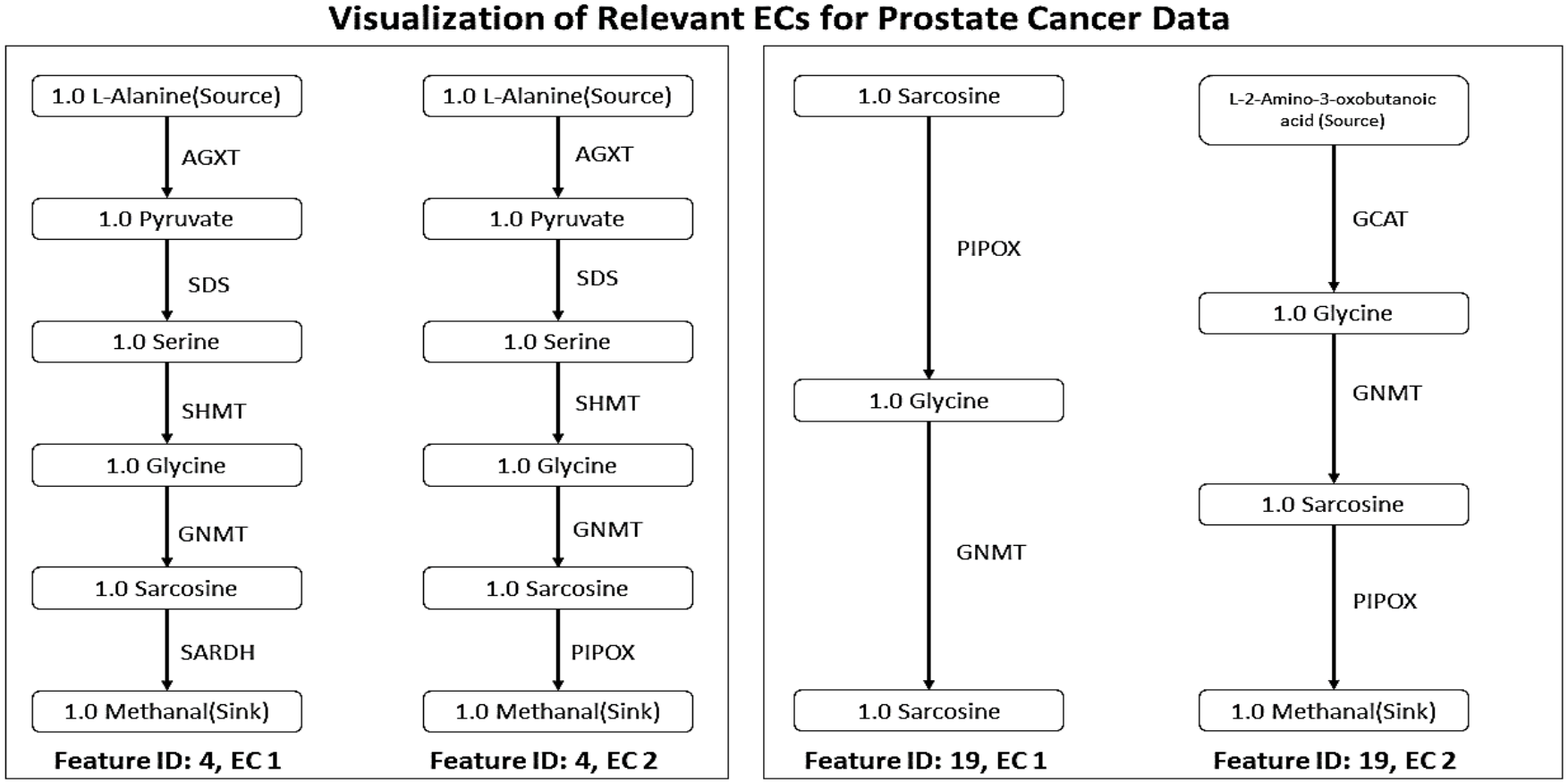


Histogram of bootstrap frequencies of 33 network features in Glycine, serine and threonine metabolism.

Pathway Name	Species/Reactions	ECs	Features	Relevant Features	Drug targets
Glycine, serine and threonine metabolism	158/55	150	33	12	31

All those features with probability >= 0.8 were declared as significant.

Results: Prostate Cancer

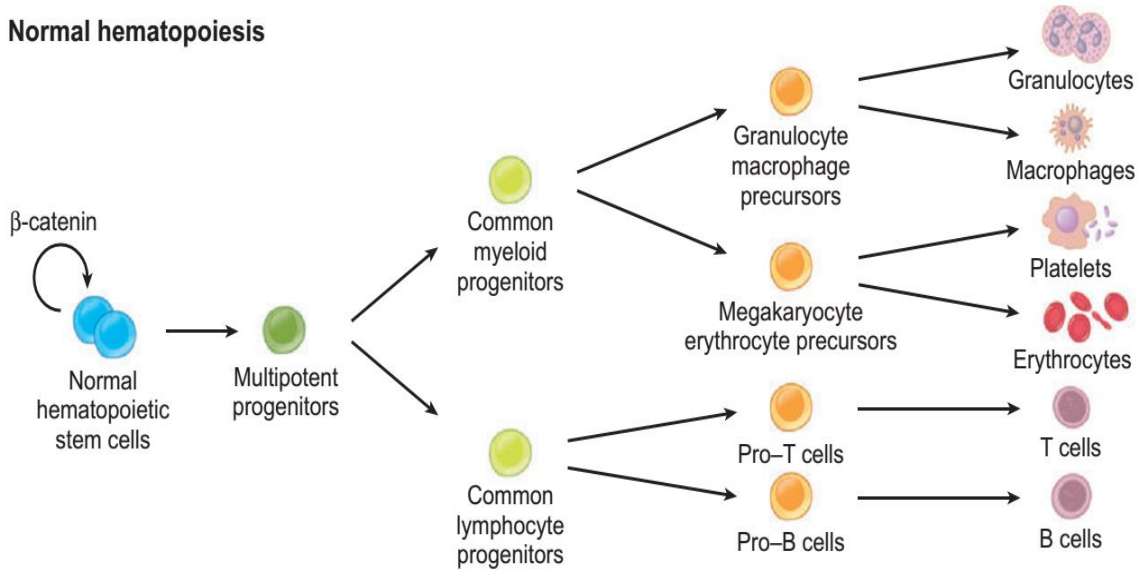


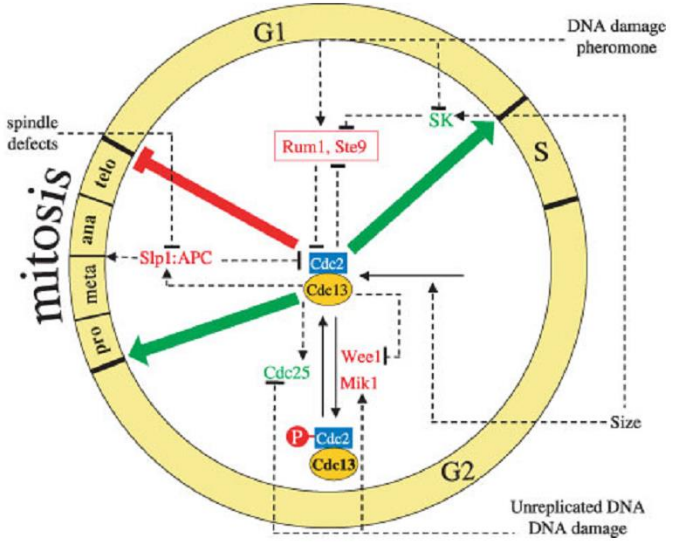
Literature Validation

1. GNMT has been associated with elevated Sarcosine levels in (Sreekumar et al., 2009) and prostate cancer progression in general (Song et al., 2011).
2. In left Sarcosine can be converted into Methanal (formaldehyde) via SARDH and PIPOX. Prostate cancer patients show increased formaldehyde concentrations in their urine (Španěl et al., 1999).

1. Motivation: Metastable States

- Biology is often understood as sequence of *biologically interpretable states*.
- Such states can be thought of being *slow regions*.

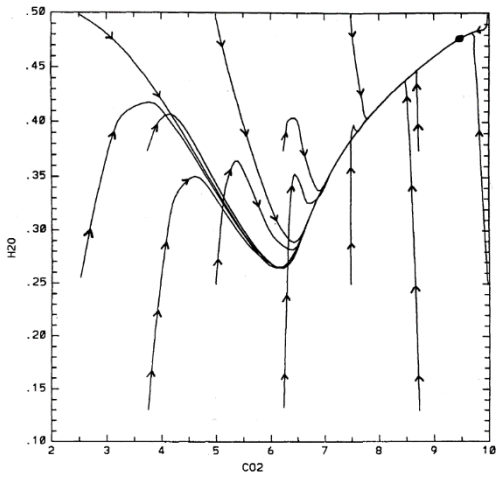


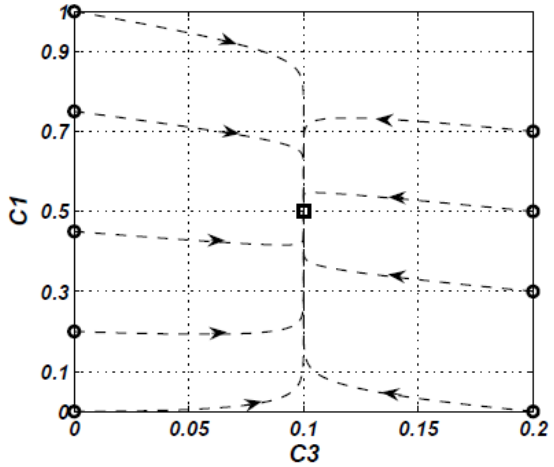


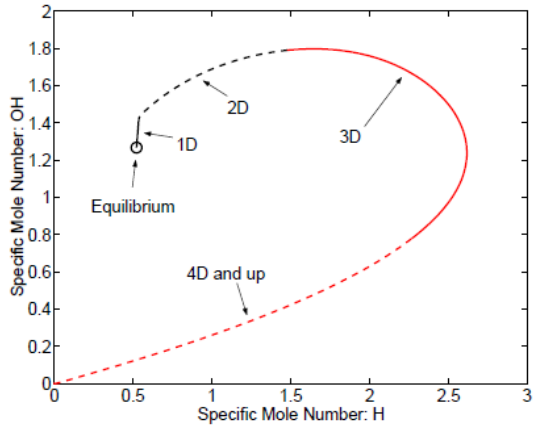
Hematopoesis from Lobo, Neethan A., et al.(2007) and Cell cycle in fission yeast from Tyson, John J., et al.(2002)

2. Motivation: Low-Dimensional Sub-Manifold

- System of ODEs often model biological processes e.g. metabolism, signalling.
- Many times, asymptotic behaviour of such systems evolve on a low-dimensional submanifold of the phase space (slow regions).

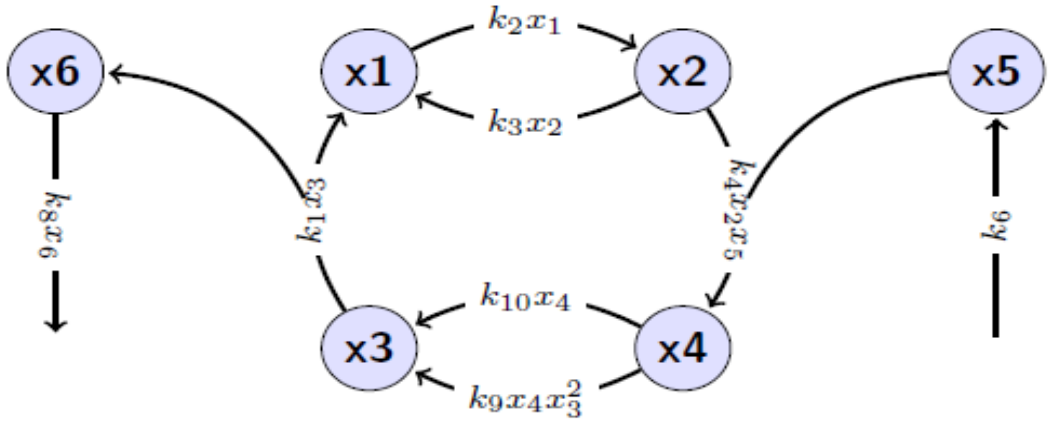






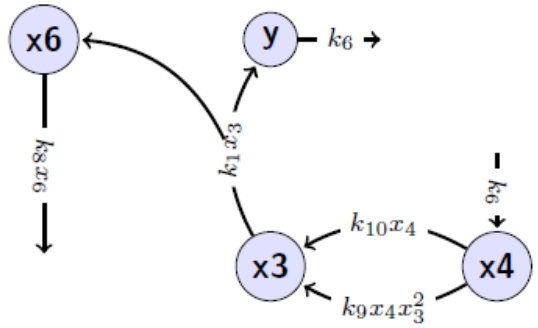
Maas, Ulrich et al.(1992), Chiavazzo, Eliodoro et al.(2007), Hung, Patrick et al.(2002)

Model Reduction: Identify Slow Variables



$$\dot{x}_1 = k_1x_3 - k_2x_1 + k_3x_2,\ \dot{x}_2 = k_2x_1 - k_3x_2 - k_4x_2x_5,$$
$$\dot{x}_3 = k_{10}x_4 - k_1x_3 + k_9x_3^2x_4,$$
$$\dot{x}_4 = k_4x_2x_5 - k_{10}x_4 - k_9x_3^2x_4,\ \dot{x}_5 = k_6 - k_4x_2x_5,$$
$$\dot{x}_6 = k_1x_3 - k_8x_6,\ x_1 + x_2 + x_3 + x_4 = 1.$$

Full Model



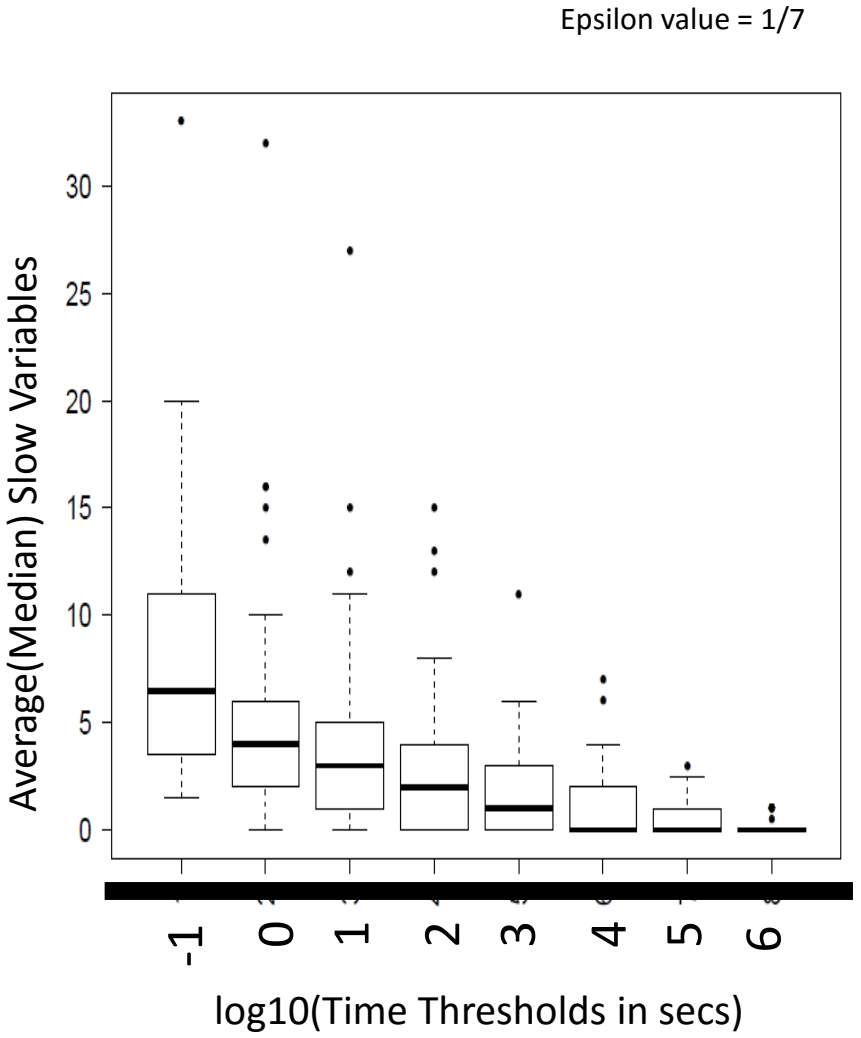
$$\dot{x}_3 = k_{10}x_4 - k_1x_3 + k_9x_3^2x_4,$$
$$\dot{x}_4 = -k_{10}x_4 + k_6 - k_9x_3^2x_4,$$
$$\dot{x}_6 = k_1x_3 - k_8x_6, \dot{y} = k_1x_3 - k_6.$$

Reduced model corresponding to one metastable state

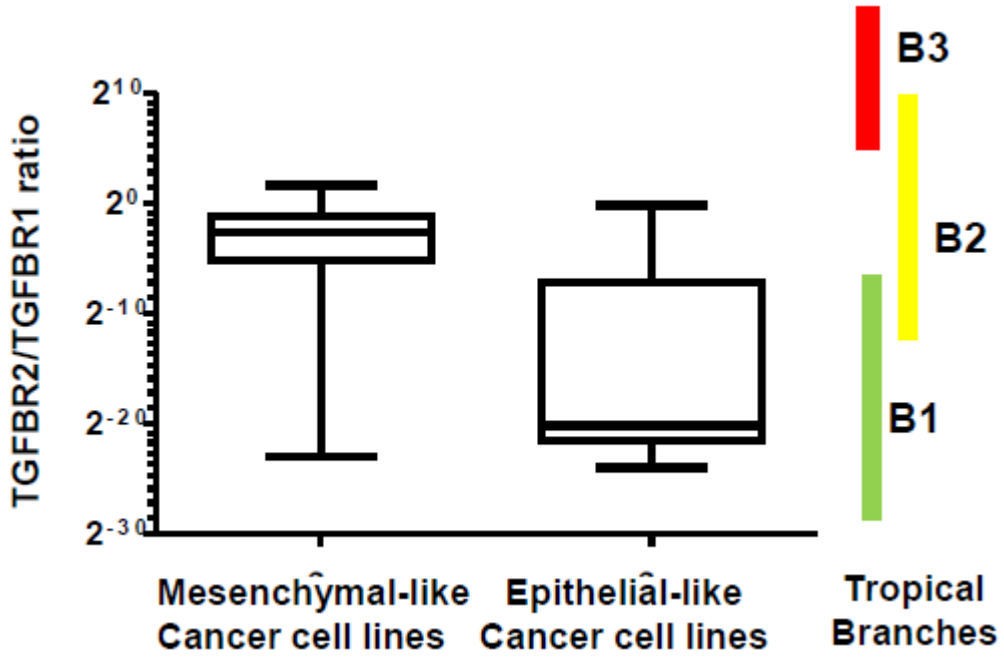
Tyson Model (cell cycle)

Benchmarking: Slow Variables

- Boxplots showing average (median) slow variables in Biomodels database for different values of time threshold.
- In this plot, a point represents the average (median) number of slow variables over all the tropical equilibrations for each model with respect to different time thresholds.
- At timescales of 1000s (in model time) and larger, reduced models have median numbers of 2 variables.

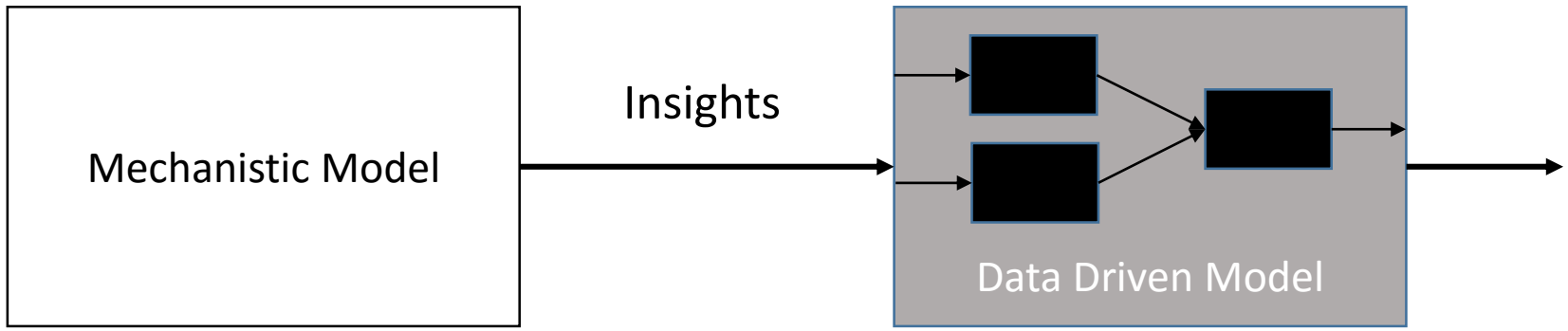


Associating States of TGFβ Model with Clinical Data



The protein expression levels of TGFBR1 and TGFBR2 was obtained from global proteome analysis of NCI-60 panel (Gholami et al. 2013)

Future Directions



Conclusion

- **Pathway-based analysis (Mostly data driven approach)**
  - Method to associate features (namely reaction pathways based on ECs) of a metabolic network to clinical or biological phenotypes with the help of gene expression data.
  - Combining data driven with mechanistic information.

- **Model Reduction (Mostly mechanistic approach)**
  - Identification of fast-slow chemical species without trajectory simulation.
  - Benchmarking on Biomodels database.
  - Very few slow variables and minimal branches.