ROBUST ALGORITHMS FOR INFERRING REGULATORY NETWORKS BASED ON GENE EXPRESSION MEASUREMENTS AND BIOLOGICAL PRIOR INFORMATION

ir. Tim Van den Bulcke

Chairman:
Prof. dr. ir. J. Berlamont

Promotors:
Prof. dr. ir. Bart De Moor
Prof. dr. ir. Kathleen Marchal

Jury:
Prof. dr. ir. J.A.K. Suykens
Prof. dr. L. De Raedt
Prof. dr. G. Verbeke
Prof. dr. ir. T. De Bie (University of Bristol (U.K.)
Dr. T. Michoel (Universiteit Gent)

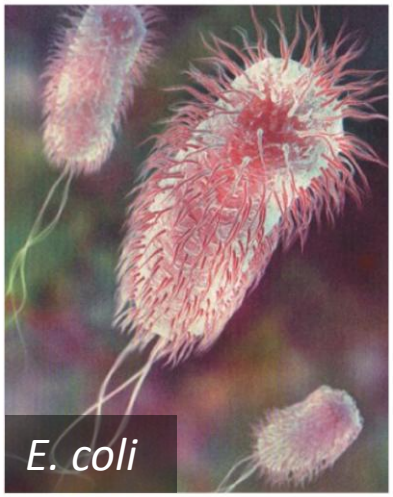
Overview

- **Overview**
- **Introduction**
  - The language of life
  - Systems biology
- **SynTReN: large scale application of simulated data to assess network inference algorithms**
  - SynTReN model
    - Network topology
  - Results
    - Effect of network topology
    - Effect of number of microarrays
- ***Pro*Bic: model-based biclustering of gene expression data**
  - *Pro*Bic model
    - Probabilistic relational models
    - EM algorithm
  - Results
    - Simulated datasets
    - *E. coli* compendium: query-driven biclustering
  - Extensions
- **Conclusion**

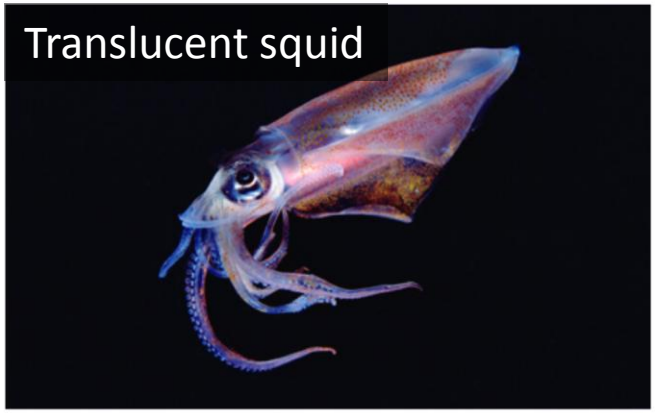
Introduction: the language of life

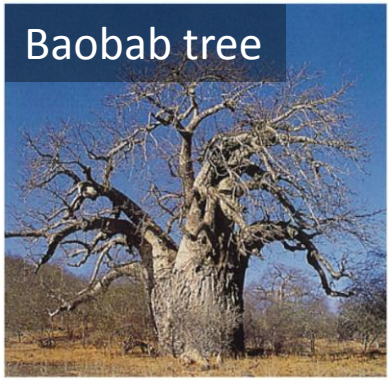
The language of life

BIOInformatics


E. coli


Meller's Chameleon


Translucent squid


Baobab tree


Humpback whale


Hippocampus


Termite mound


Coral snake

The language of life: building blocks







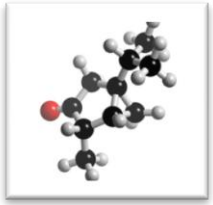

The language of life: building blocks

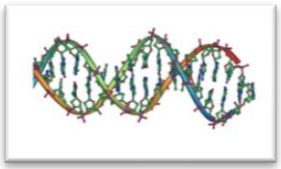
RNA
messenger

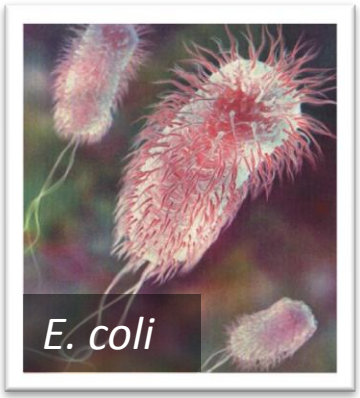
Protein
workhorses

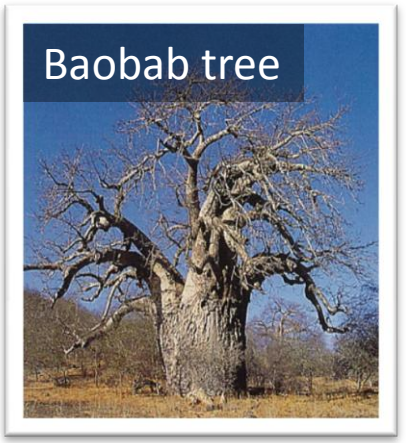
Metabolite
signaling, activation

DNA / genes
information carrier









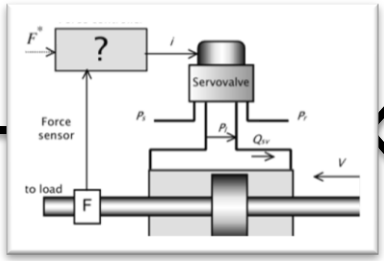
The language of life: regulation

Steering action

Steering

Signal transduction

Signal conversion to hydraulic pressure

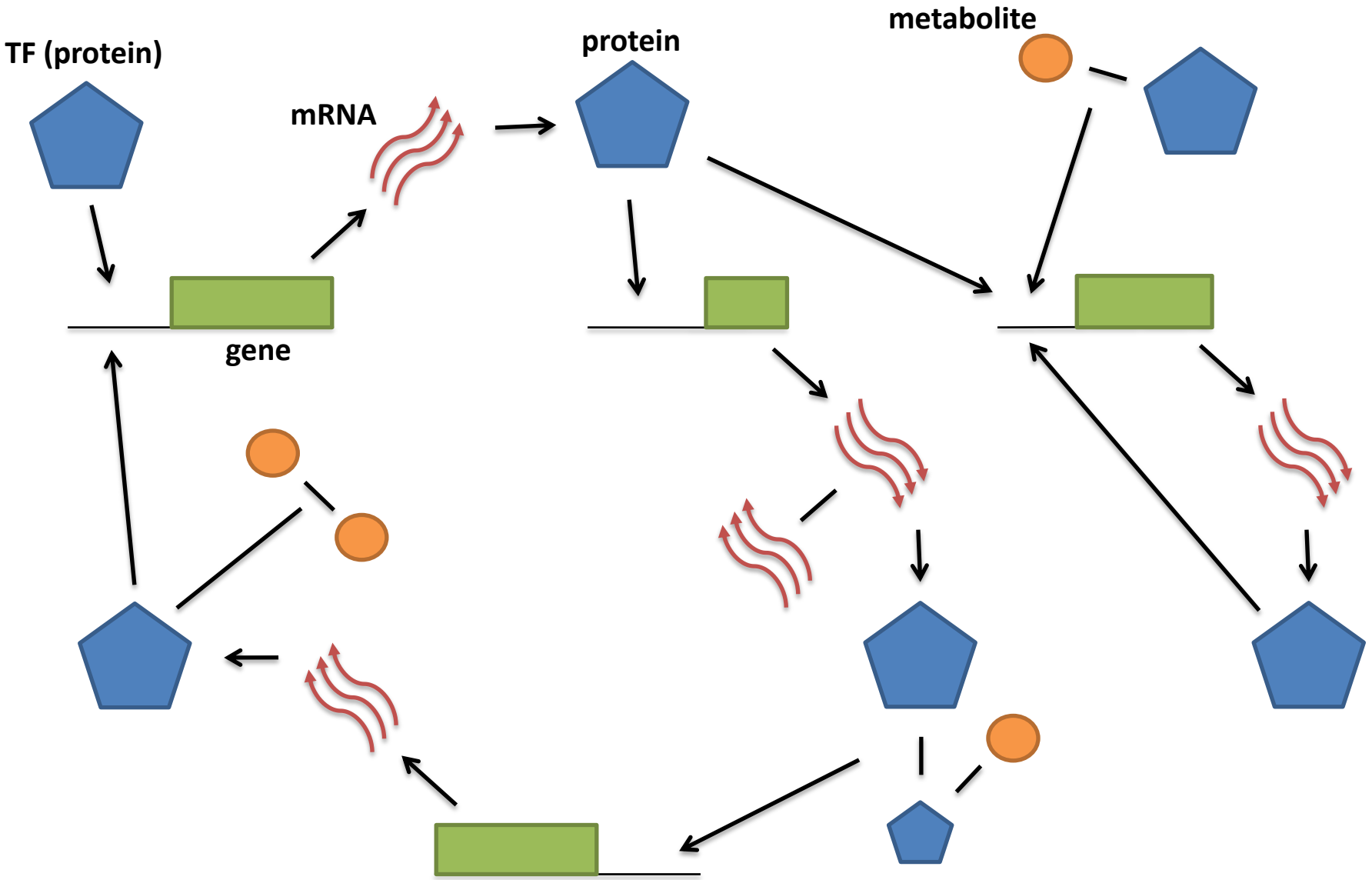


Flap movement

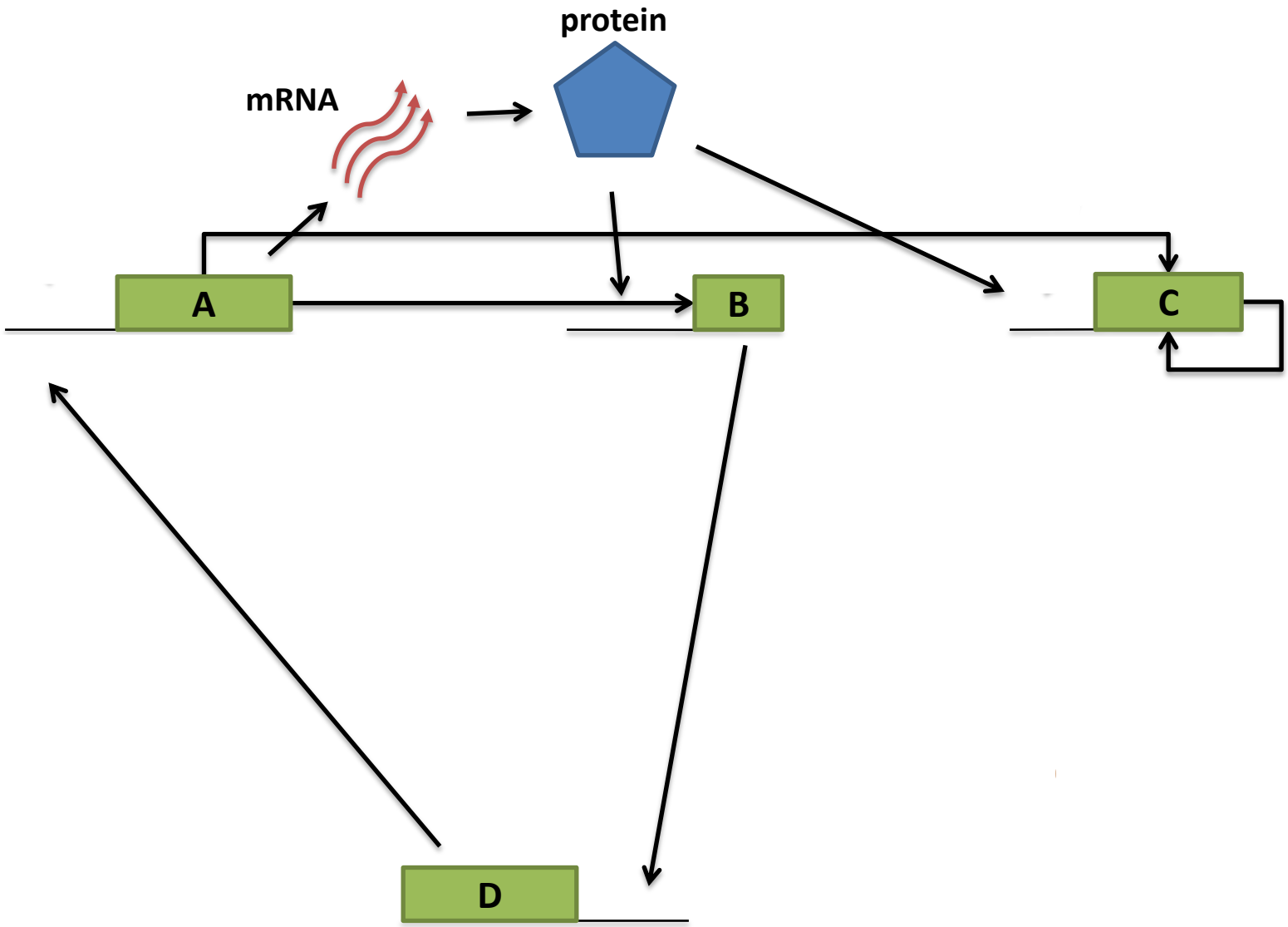
Black box recording

Feedback information

The language of life: regulation

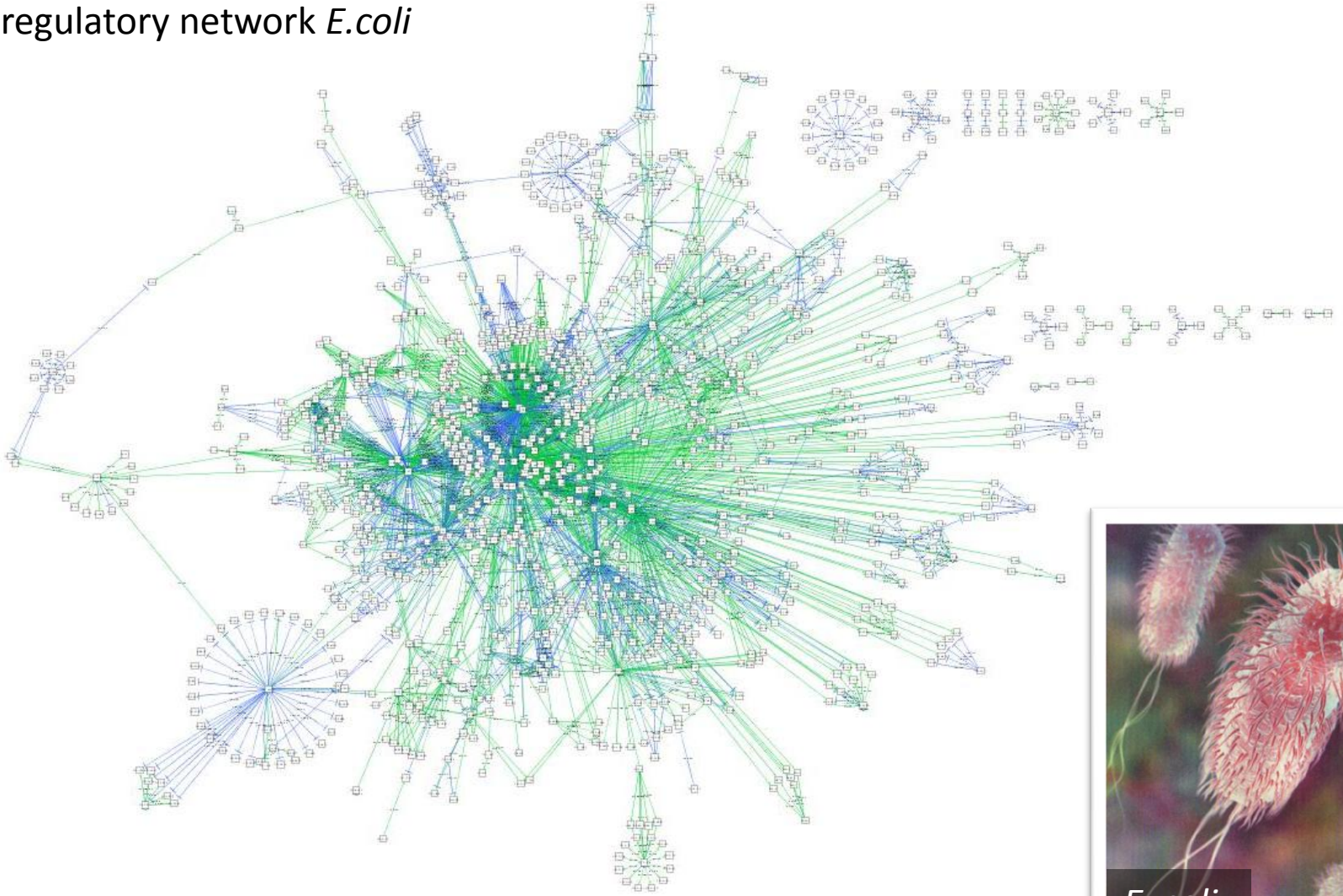


The language of life: regulation



The language of life: regulation

Gene regulatory network *E.coli*



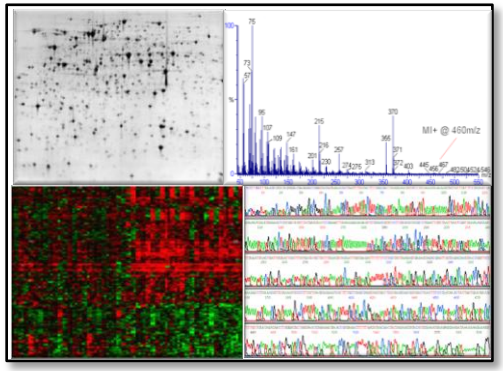
Introduction: systems biology

Systems biology

- **Systems biology**
  - Systematic study of complex **interactions** in biological systems using a **holistic** approach

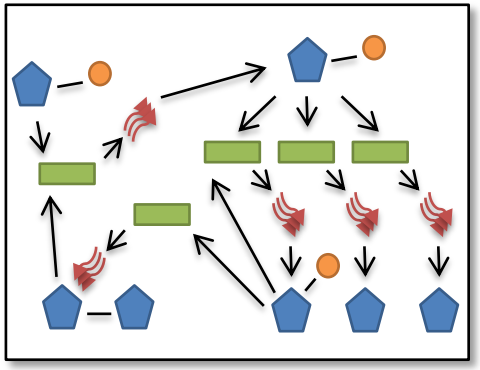
Organism

High-throughput experiments



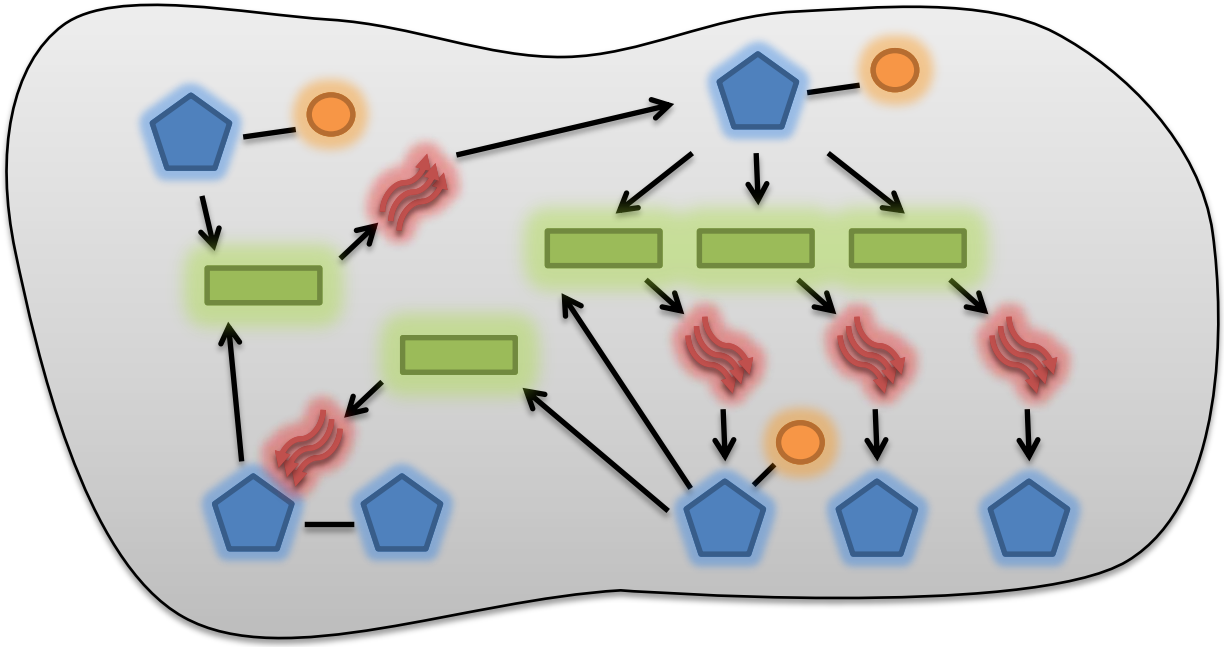
***'omics* data**

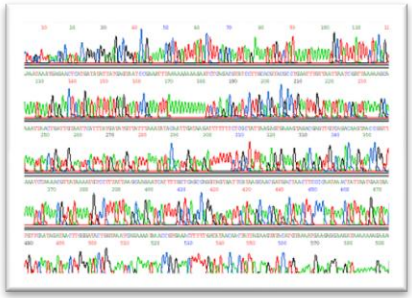
Network inference algorithm



Interaction network

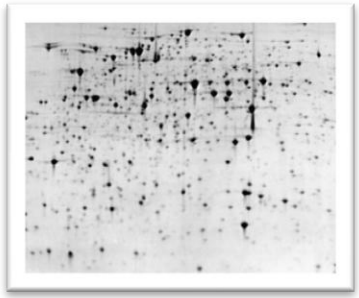
High-throughput *'omics* data





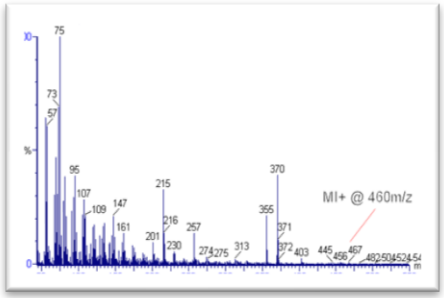
Genomics

e.g. DNA sequencing



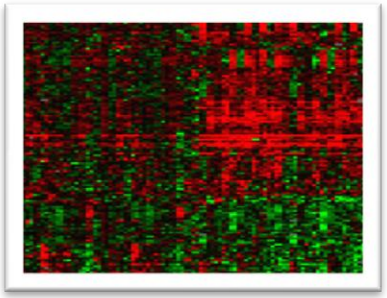
Proteomics

e.g. 2D gels



Metabolomics

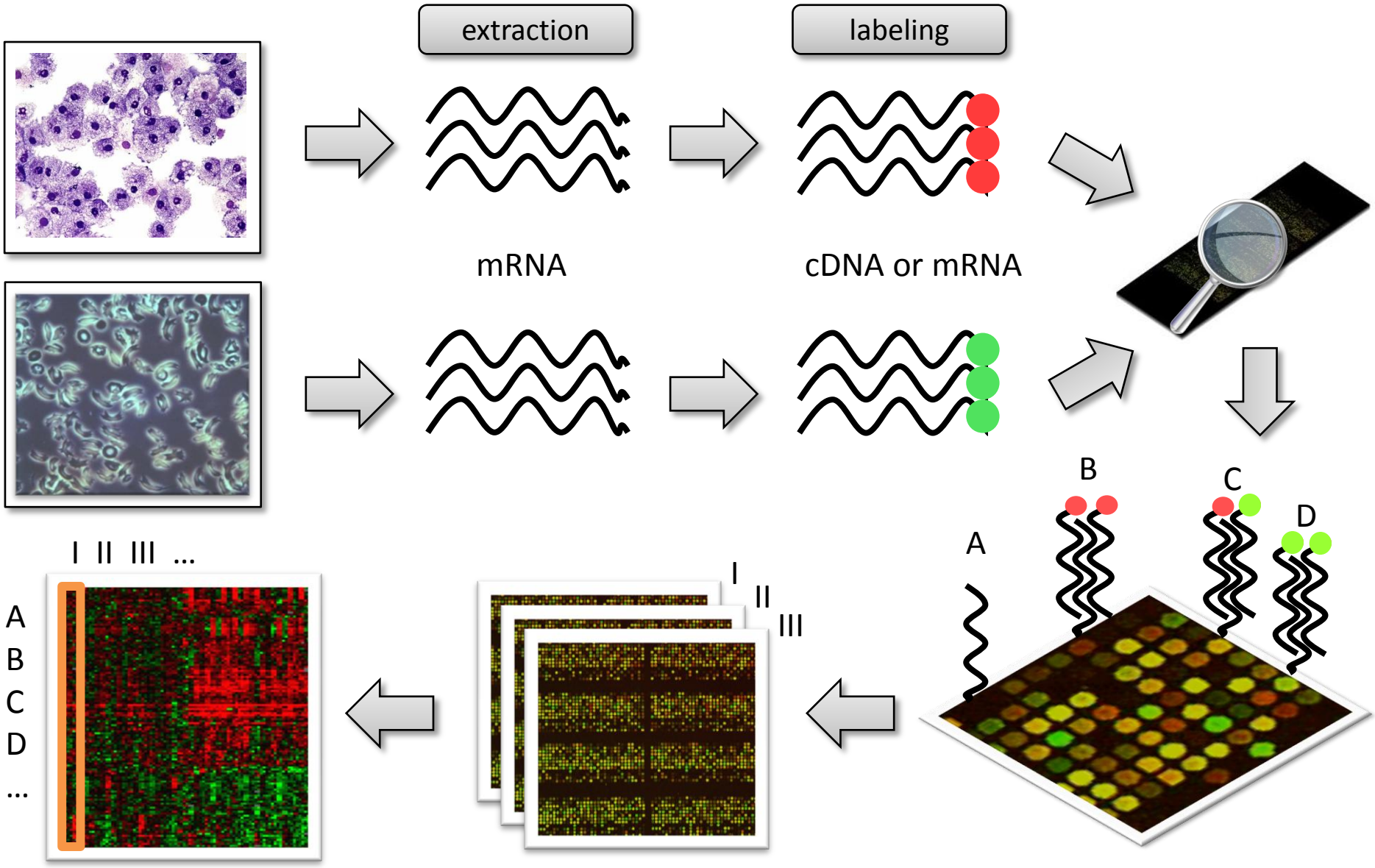
e.g. mass spectrometry



Transcriptomics

e.g. DNA microarrays

DNA microarrays



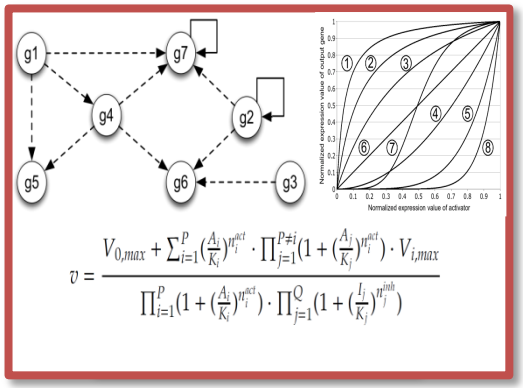
Part I: SynTReN

Large scale application of simulated data
to assess network inference algorithms

Introduction

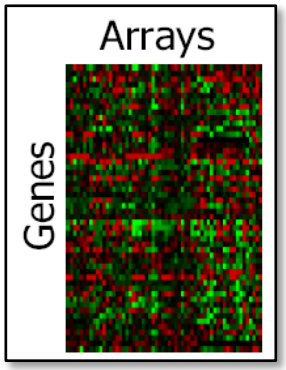
Reconstruction of gene regulatory networks

Simulated organism



Microarray experiments

Microarray data



Network inference algorithm

Gene regulatory network

Limited knowledge of true underlying network

→ How to characterize algorithm performance?

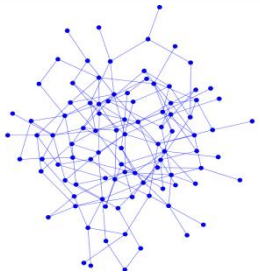
Introduction

- Simulated data
  - (+) Complete insight in underlying model (*white box* model)
  - (+) Control over all parameters
  - (+) Fast, *in silico* experiments
  - (-) Based on simplified models of biology

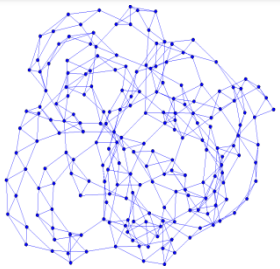
- **SynTReN**
  - Fast generation of large simulated gene expression datasets under various settings
  - Offer additional, sometimes unexpected, insights in the behavior of inference algorithms compared to biological data only

SynTReN setup

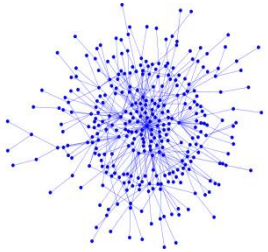
Random graph models



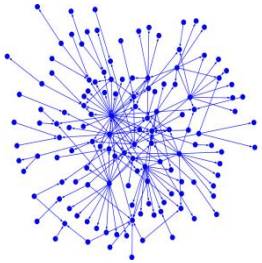
Erdös-Rényi



Watts-Strogatz

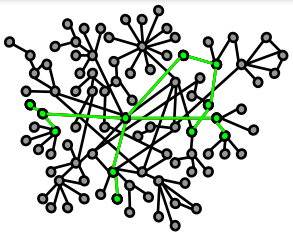


Albert-Barabási

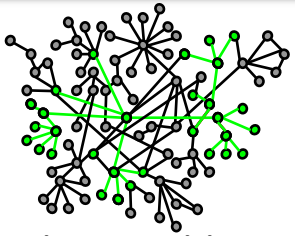


Directed scale free

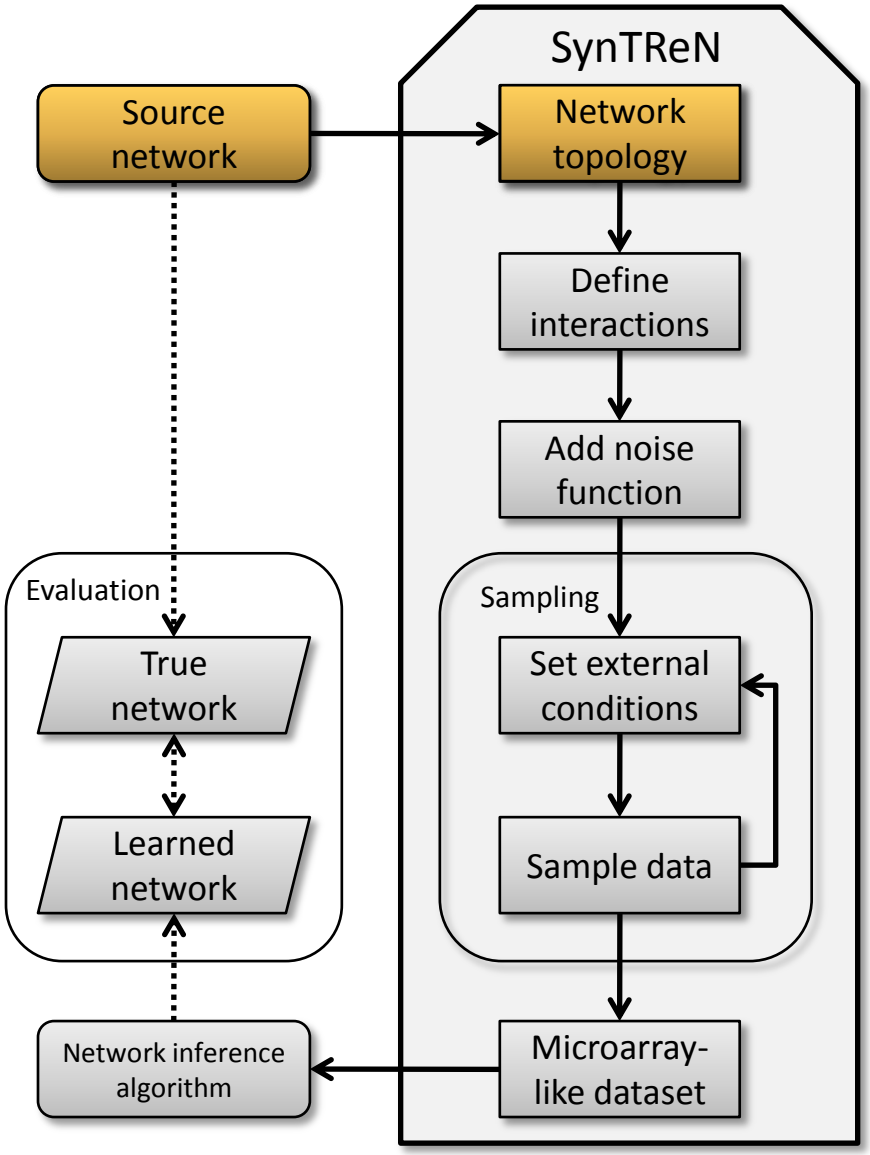
Subnetwork selection



Neighbor addition



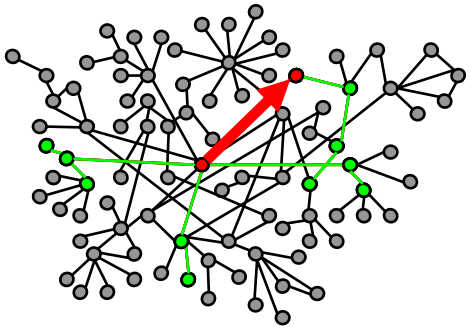
Cluster addition

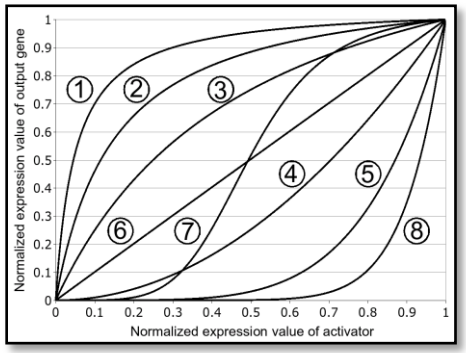


SynTReN setup

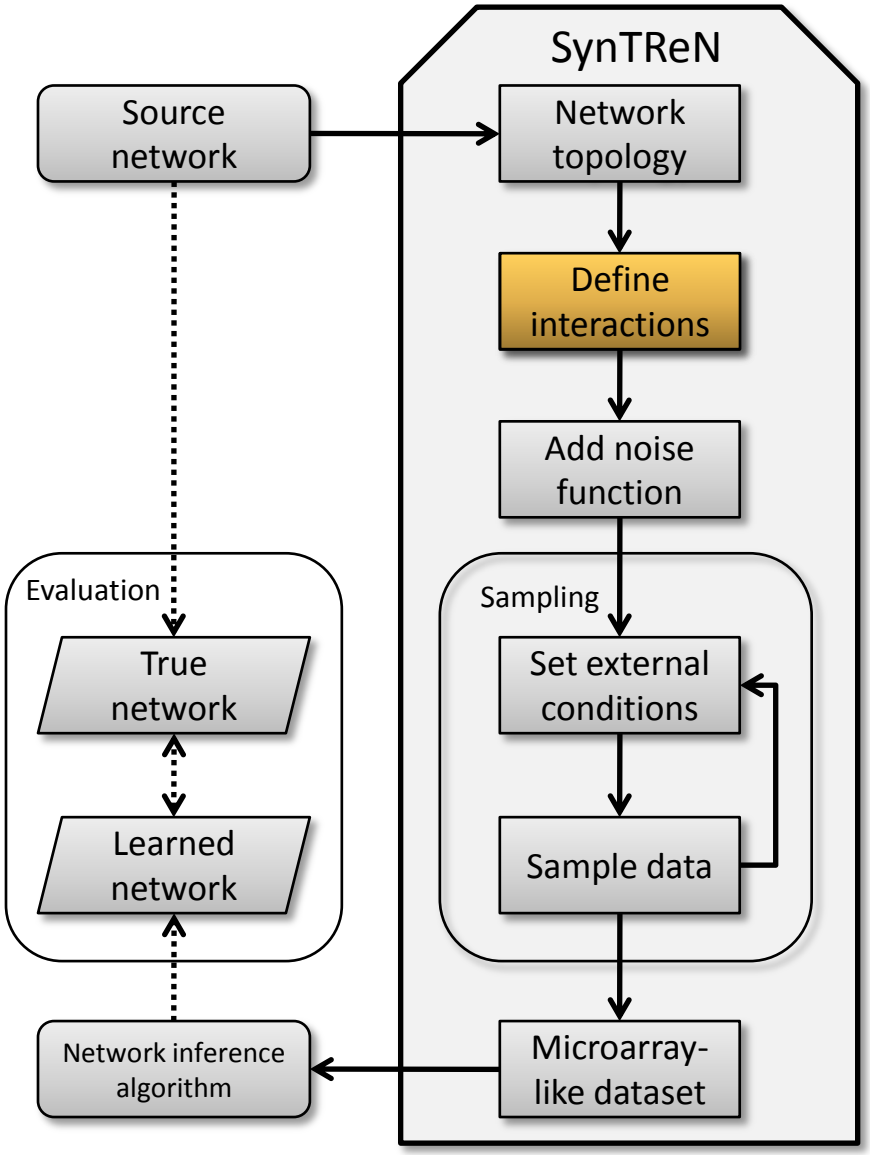
Michaelis-Menten and Hill enzyme kinetics

- Complex interactions
  - Synergism and antagonism
  - Cooperative binding
  - Competitive binding
- Steady-state





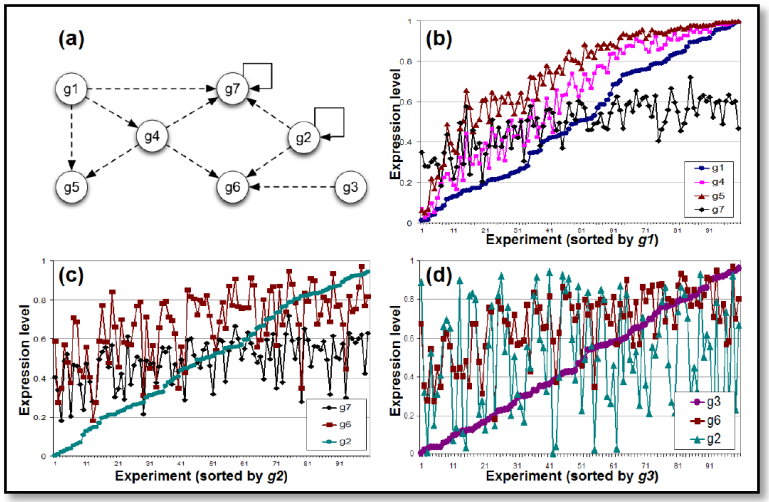
$$v = \frac{V_{0,max} + \sum_{i=1}^{P}(\frac{A_i}{K_i})^{n_i^{act}} \cdot \prod_{j=1}^{P\neq i}(1 + (\frac{A_j}{K_j})^{n_i^{act}}) \cdot V_{i,max}}{\prod_{i=1}^{P}(1 + (\frac{A_i}{K_i})^{n_i^{act}}) \cdot \prod_{j=1}^{Q}(1 + (\frac{I_j}{K_j})^{n_j^{inh}})}$$

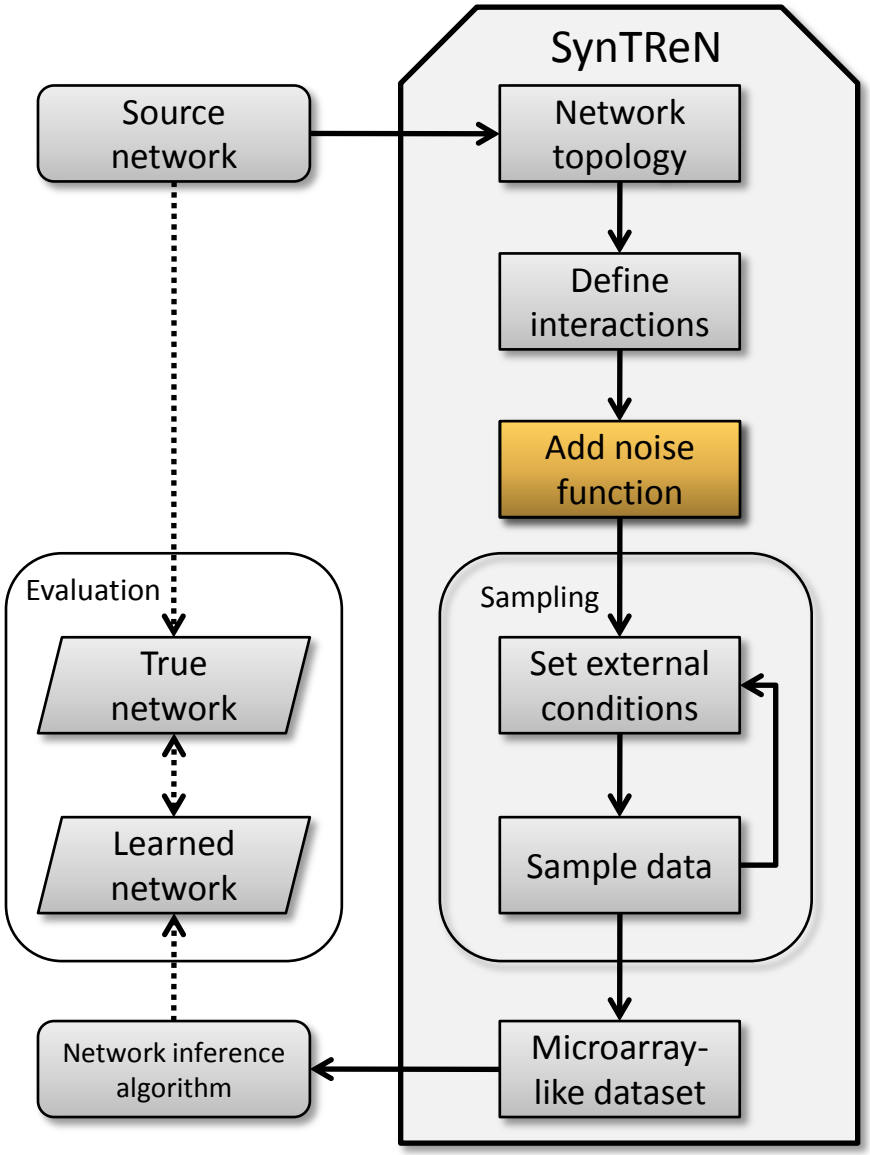


SynTReN setup

Adding different noise types

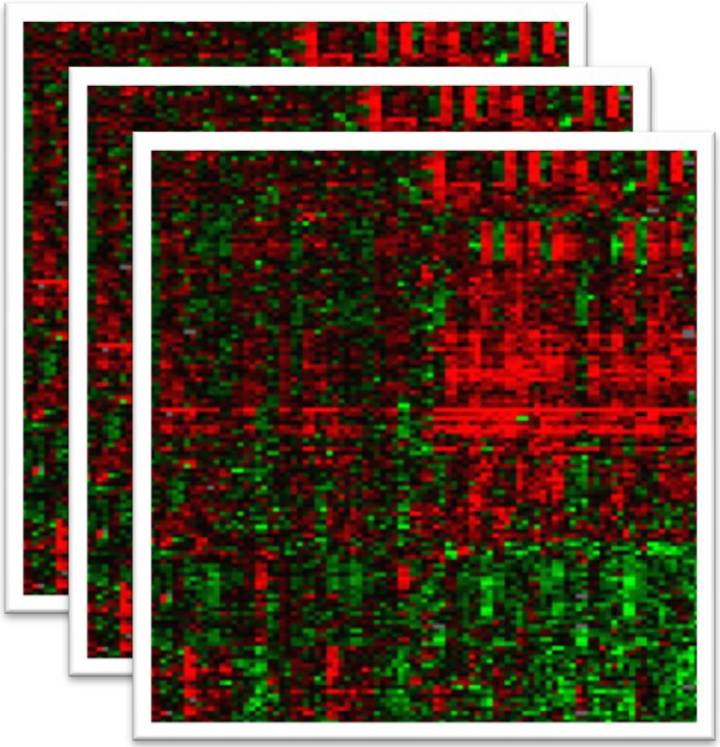
- Different noise types, caused by:
  - Experimental setup
  - Biological variation
  - External stimuli

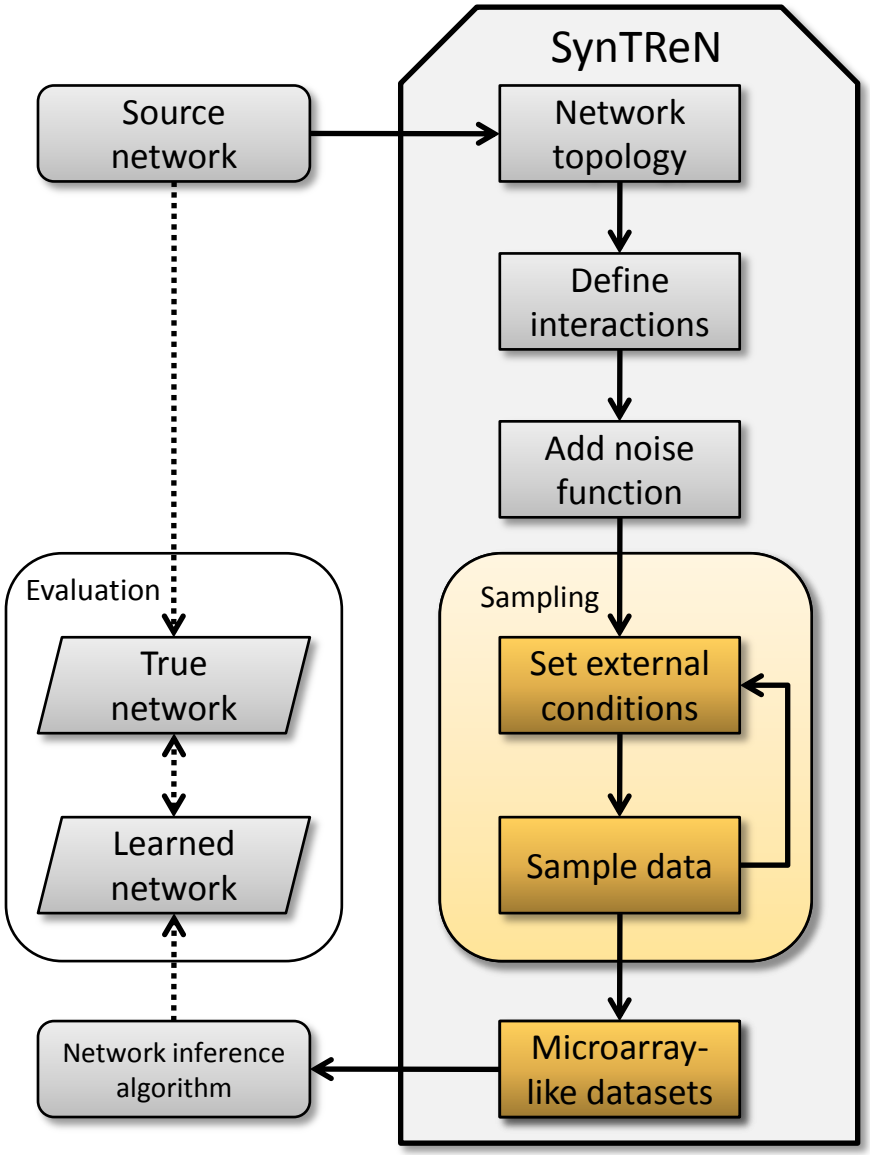




SynTReN setup

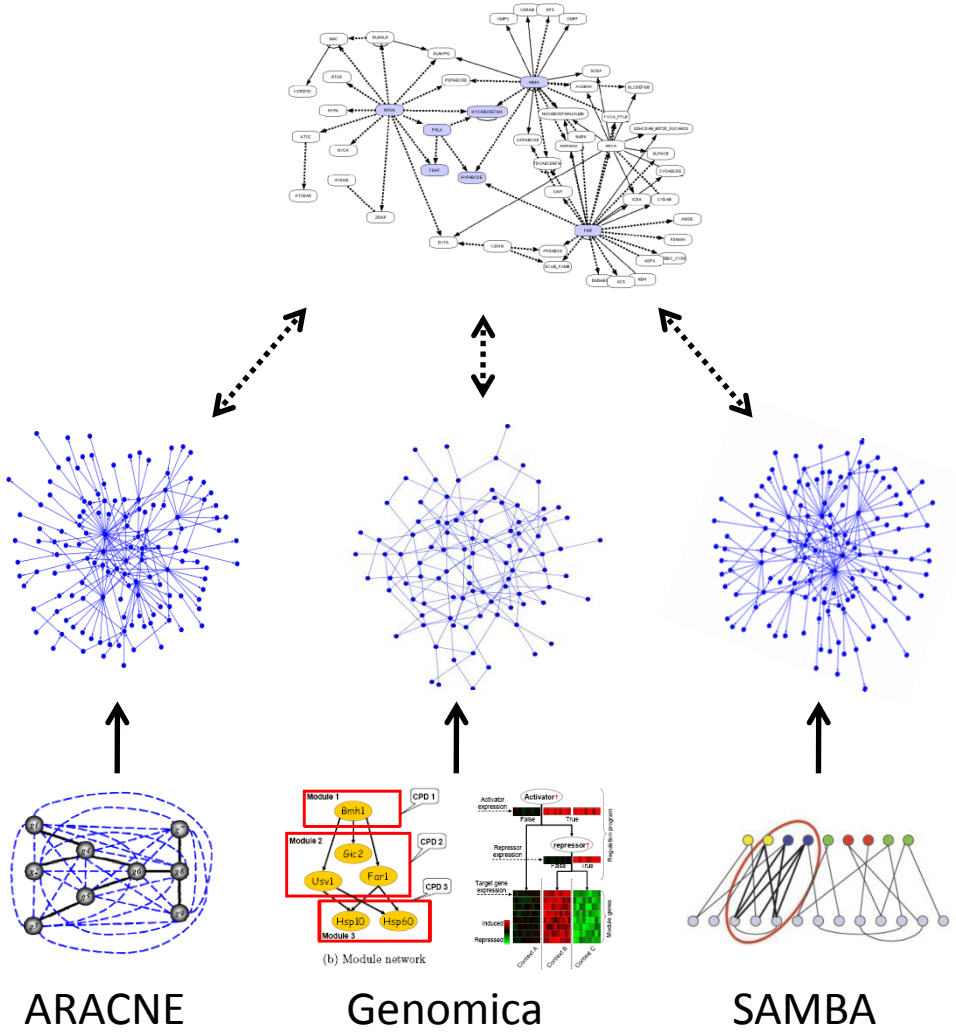
Sampling data

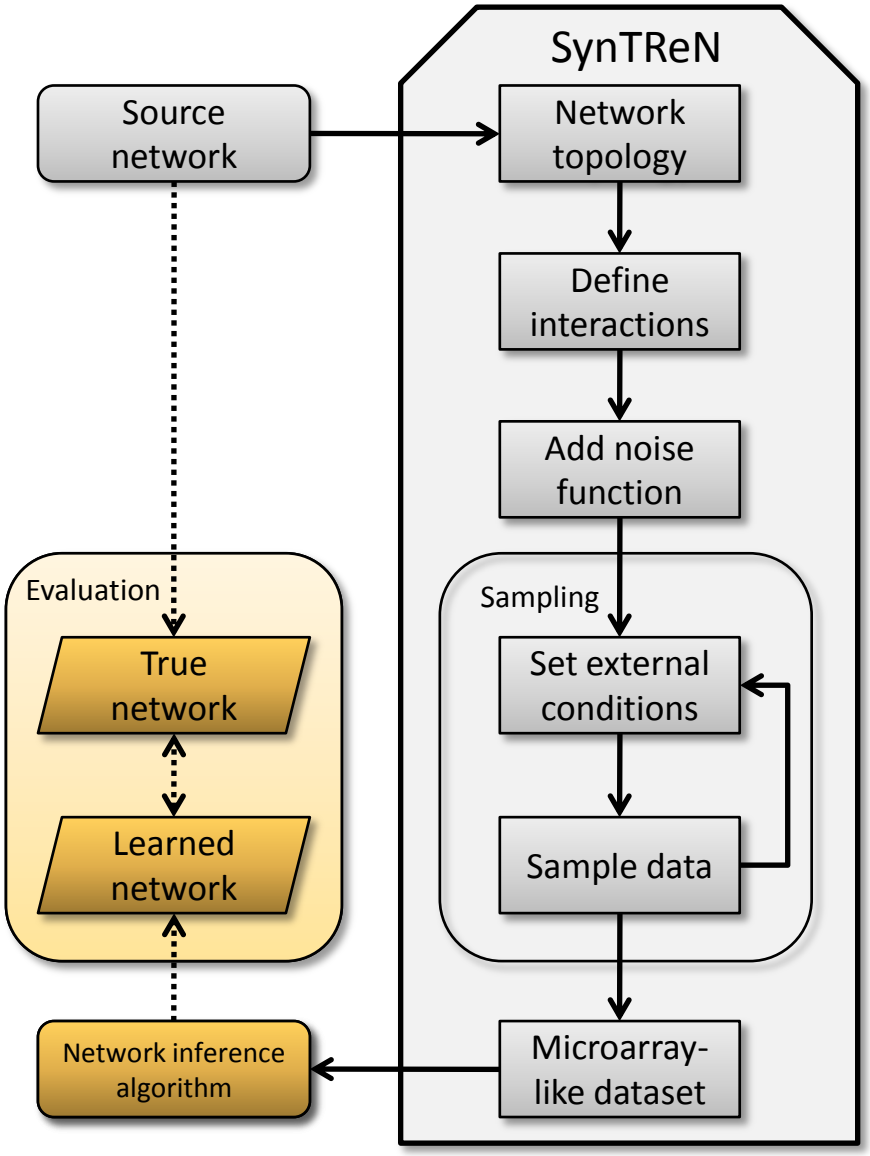




SynTReN setup

Characterization of network inference algorithms



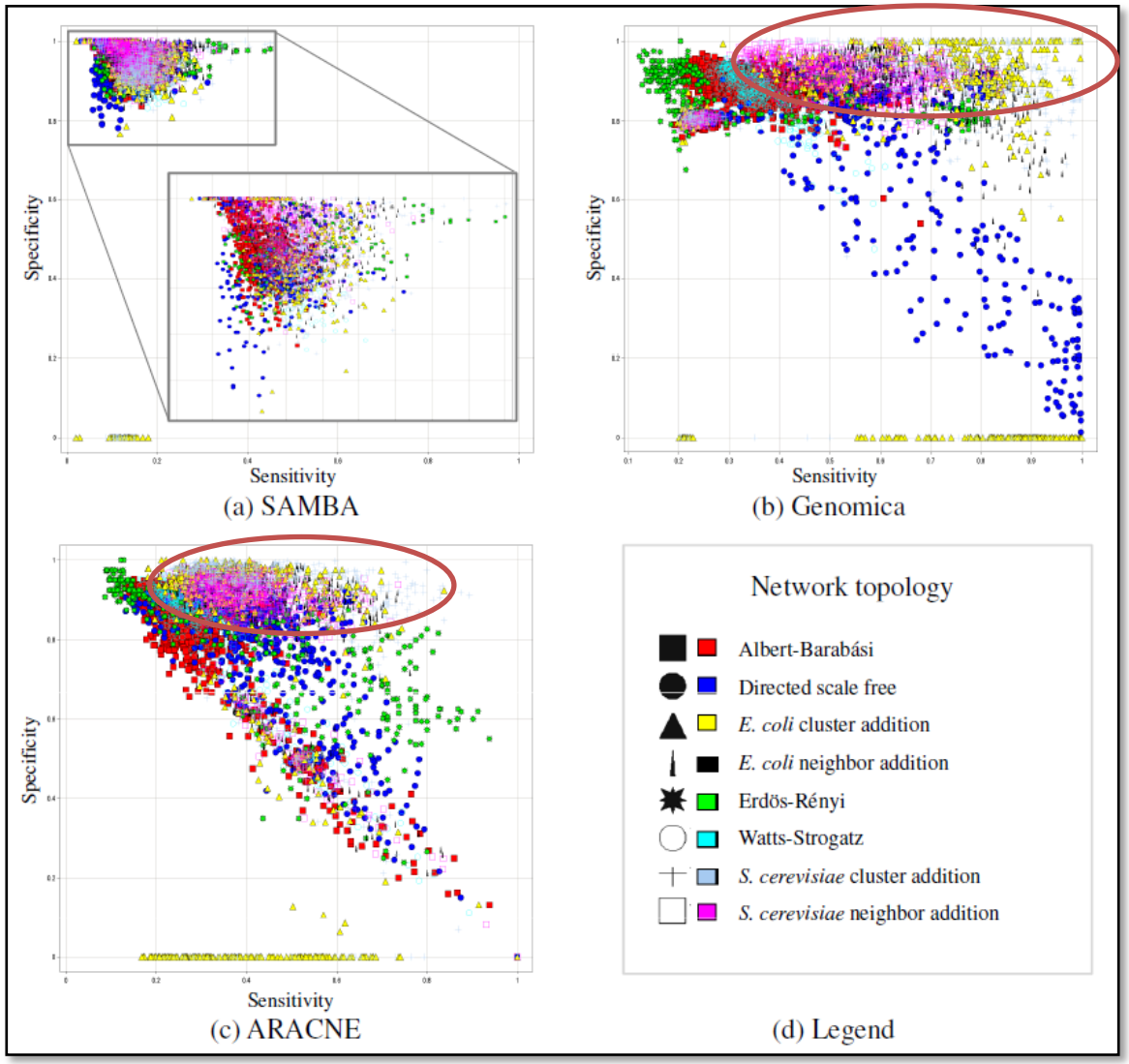


SynTReN: results

SynTReN results

- Effect of network topology
- Effect of noise and amount of data

Effect of network topology



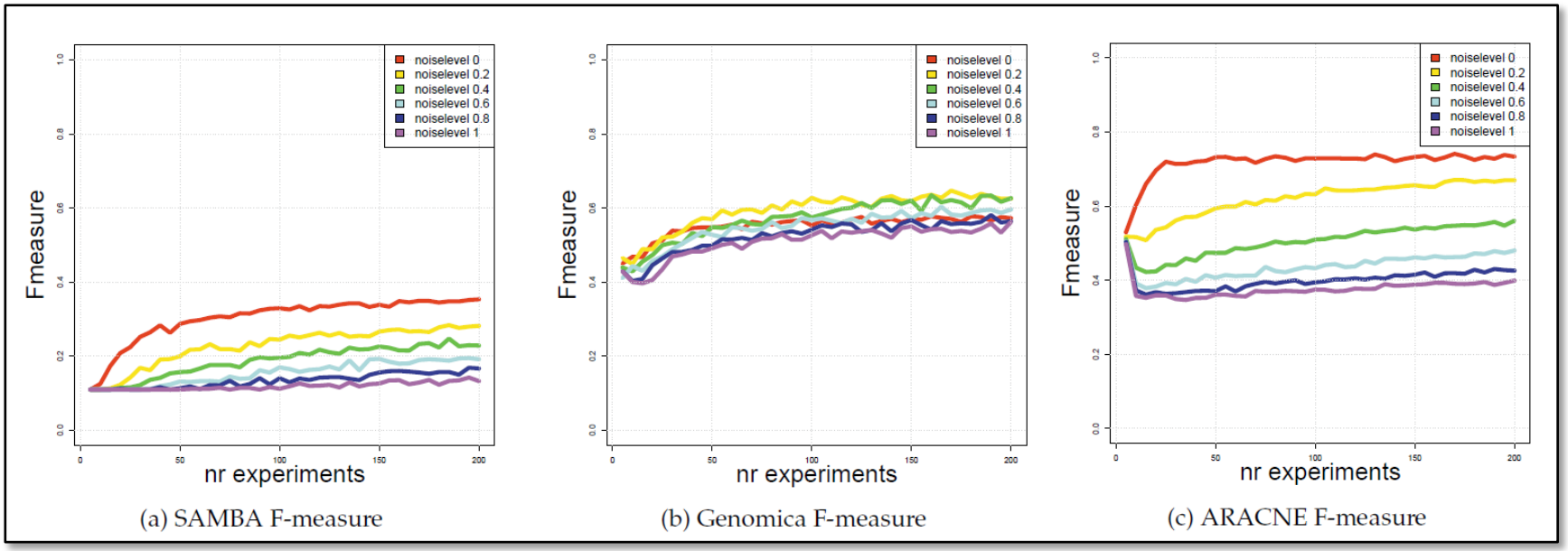
- Network topology has a strong impact on algorithm performance
- Genomica and ARACNE perform significantly better on biological (sub)networks

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

Effect of noise and amount of data

- Algorithms show qualitatively different response to increasing data
- Plateau is reached for some algorithms
- Varying dependency on noise



SynTReN conclusions

- Application simulated data offers interesting insights in characteristics of network inference algorithms
- Underlying network topology of simulated data is an important factor w.r.t. the quality of the inferred network
- Biological (sub)networks generally lead to better inference

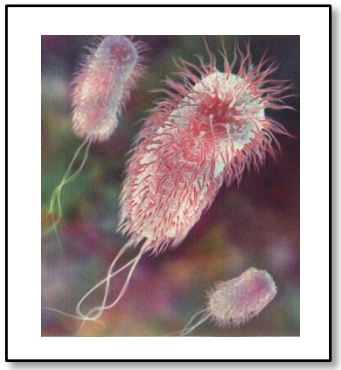
Part II: *Pro*Bic

Model-based biclustering of gene expression data

*Pro*Bic: introduction

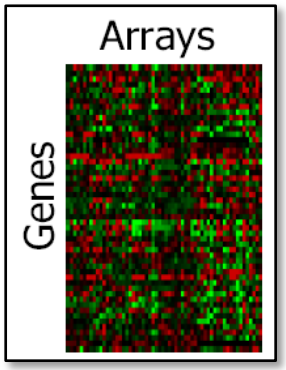
Reconstruction of gene regulatory networks

Organism



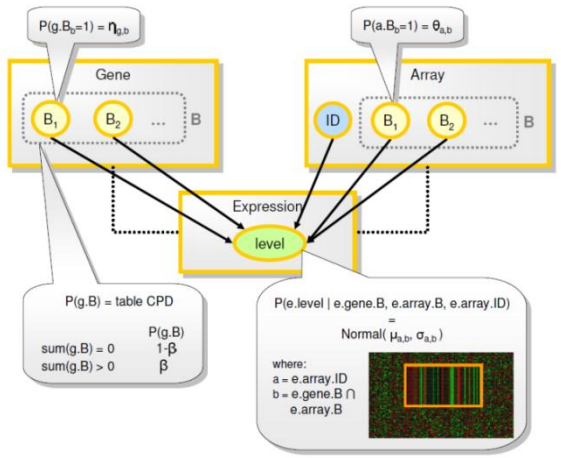
Microarray experiments

Microarray data



Network inference algorithm

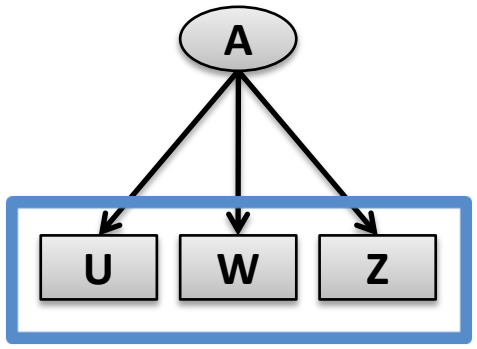
Gene regulatory network



Model-based biclustering

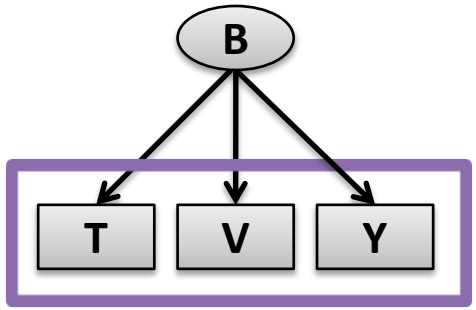
*Pro*Bic: introduction

Biclustering



1-4

Global biclustering



3-5

Query-driven biclustering

Gene	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Gene T					
Gene U					
Gene V					
Gene W					
Gene X					
Gene Y					
Gene Z					

*Pro*Bic introduction

*Pro*Bic goals:

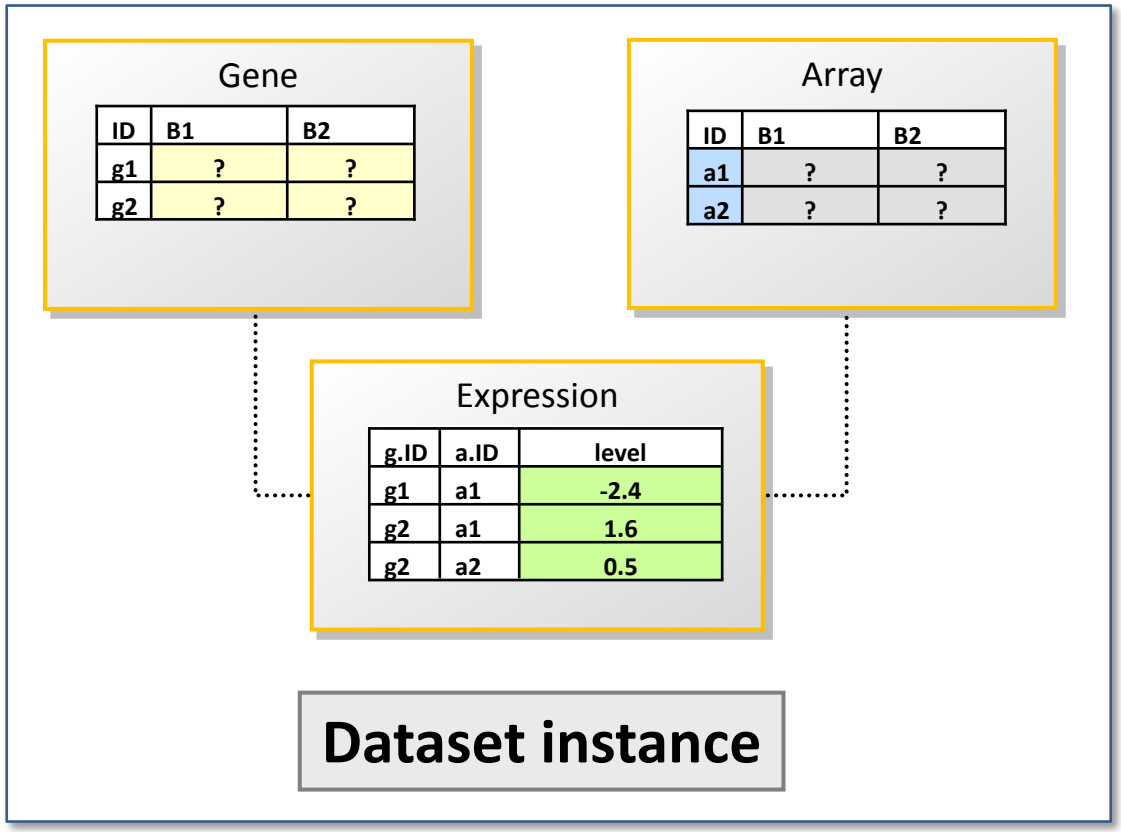
- Unified probabilistic model
- Combined query-driven and global biclustering
- Model driven:
  - Multiple overlapping biclusters
  - Incorporate diametrical biclustering
- Computational efficiency
- No data discretization required
- Extensible towards heterogeneous data

*Pro*Bic model

Based on *Probabilistic relational model* framework

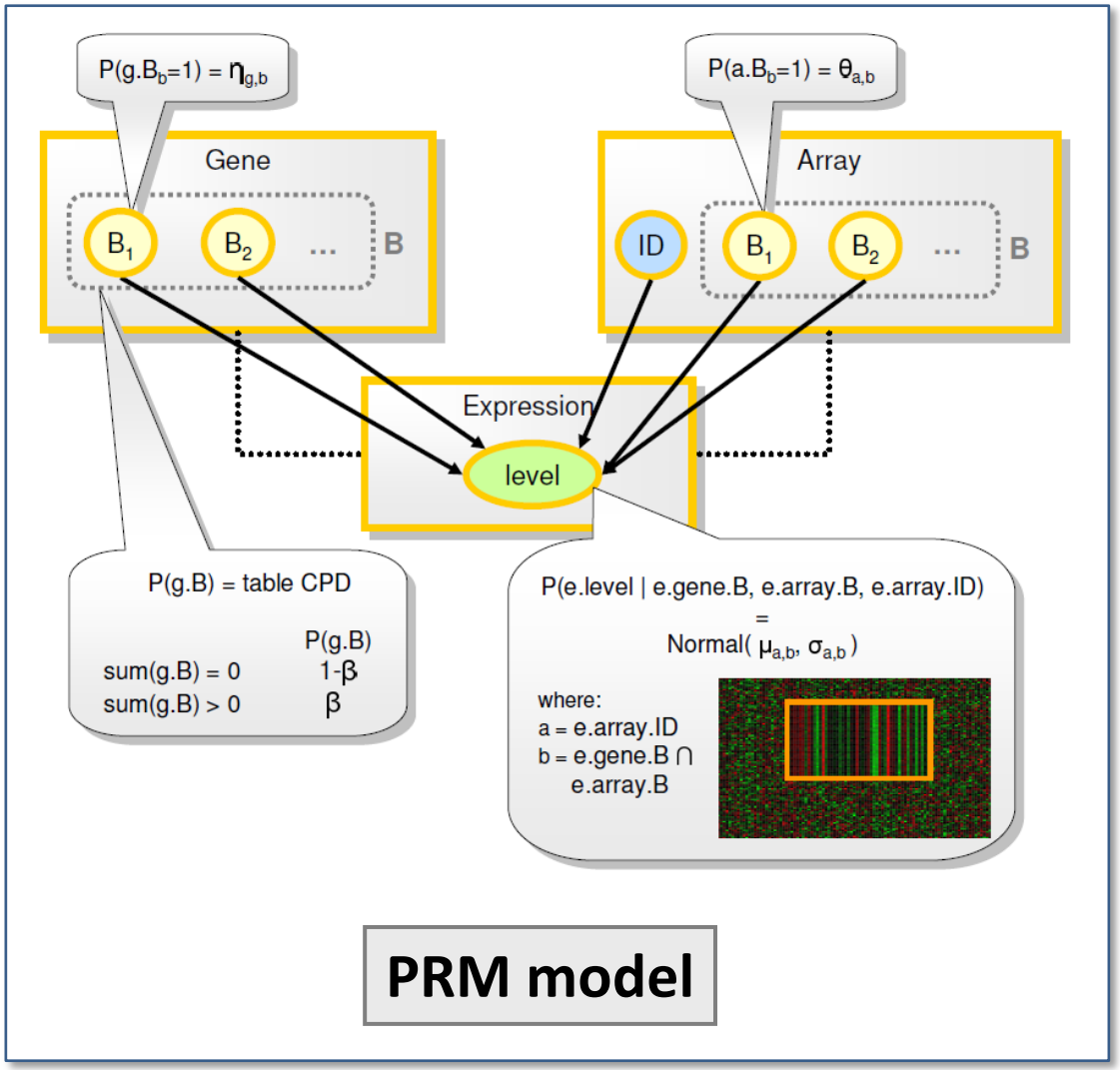
- relational extension to Bayesian networks:

Bayesian networks	~	a single flat table
Probabilistic relational models	~	relational data structure



- Designed for:
  - Computational efficiency
  - Extensibility
- Three classes:
  - *Gene*
  - *Array*
  - *Expression*

*Pro*Bic model

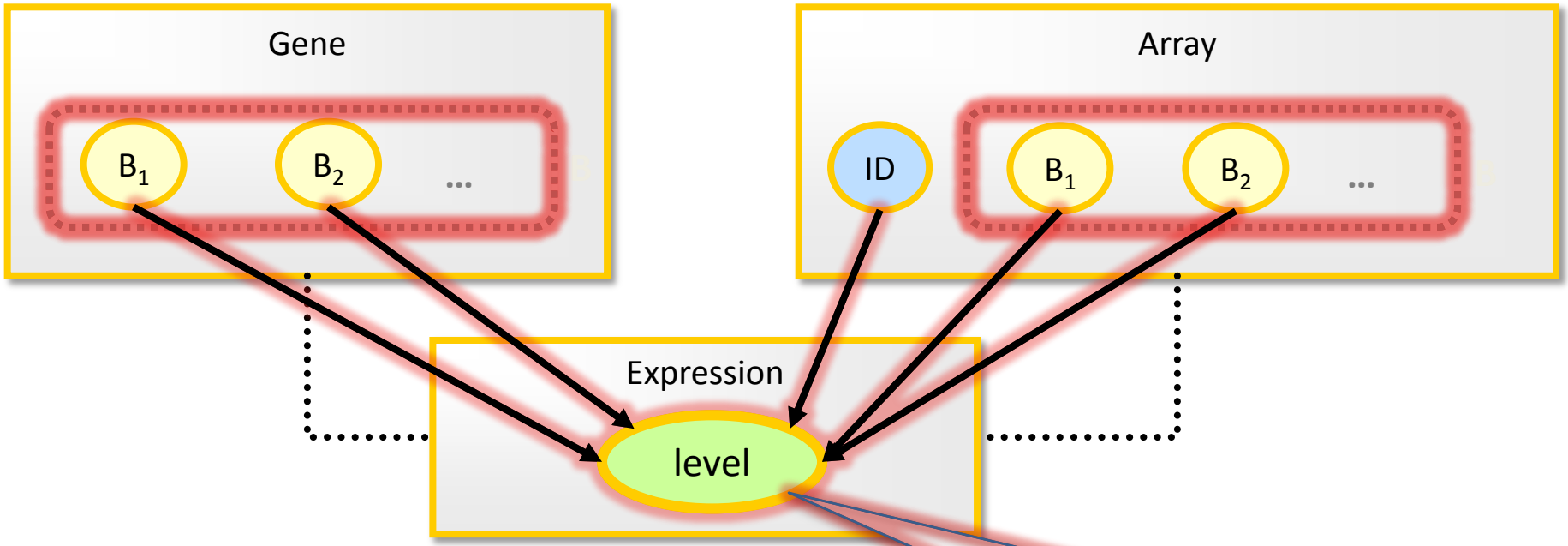


- Parameters:
  - Set of Normal distributions
- Hidden variables:
  - Gene-bicluster assignments
  - Array-bicluster assignments
- Priors:
  - Query-driven biclustering
  - Expert knowledge

ProBic model: EM algorithm

- Only approximative algorithms are tractable
- Hard assignment Expectation-Maximization algorithm
  - Natural decomposition of the model in the EM steps
  - Efficient
  - Extensible
  - Good convergence properties

ProBic model: EM algorithm



Initialization

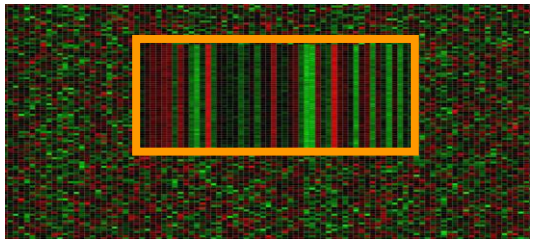
While not converged:

- Maximization step
  - Normal distribution parameters
- Expectation step
  - gene-bicluster assignments
  - array-bicluster assignments

P(e.level | e.gene.B, e.array.B, e.array.ID)

=

Normal($\mu_{a,b}$, $\sigma_{a,b}$)



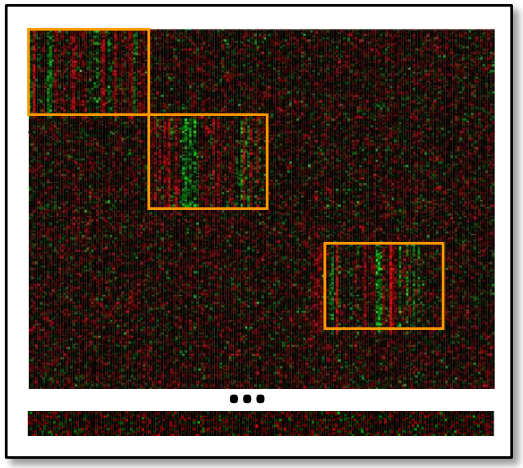
*Pro*Bic: results

*Pro*Bic results

- Simulated datasets:
  - Noise and missing value robustness
  - Comparison to state of the art
- Query-driven biclustering (*E. coli* compendium):
  - Single gene queries
  - Outlier removal in multi-gene queries

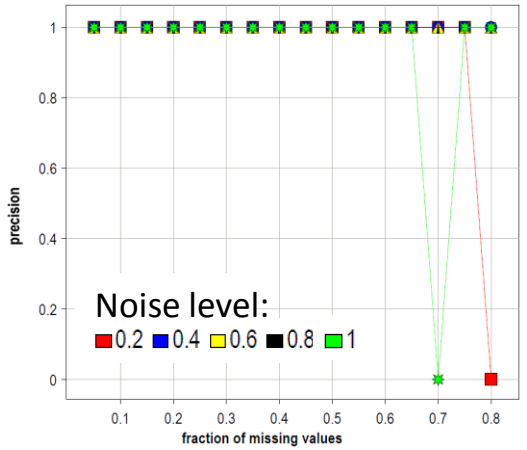
Noise and missing value robustness

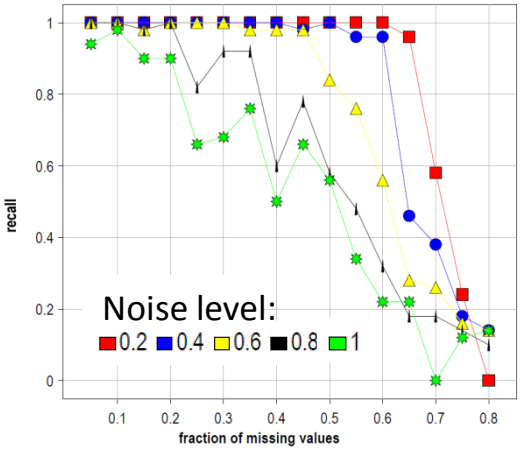
- Simulated data: 500x200
- 3 biclusters (50x50)
- Varying noise and missing values

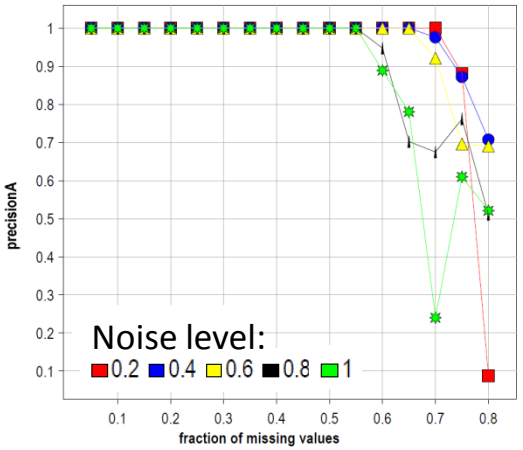


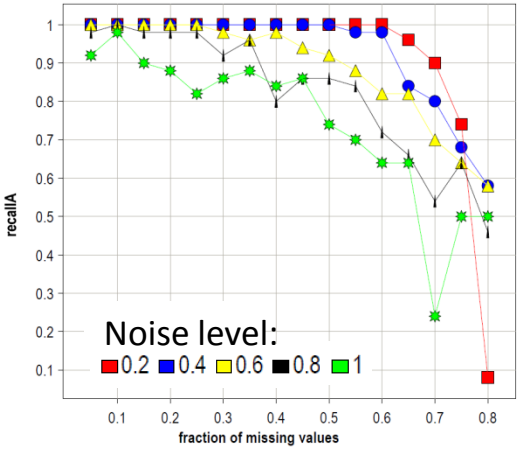
genes

arrays









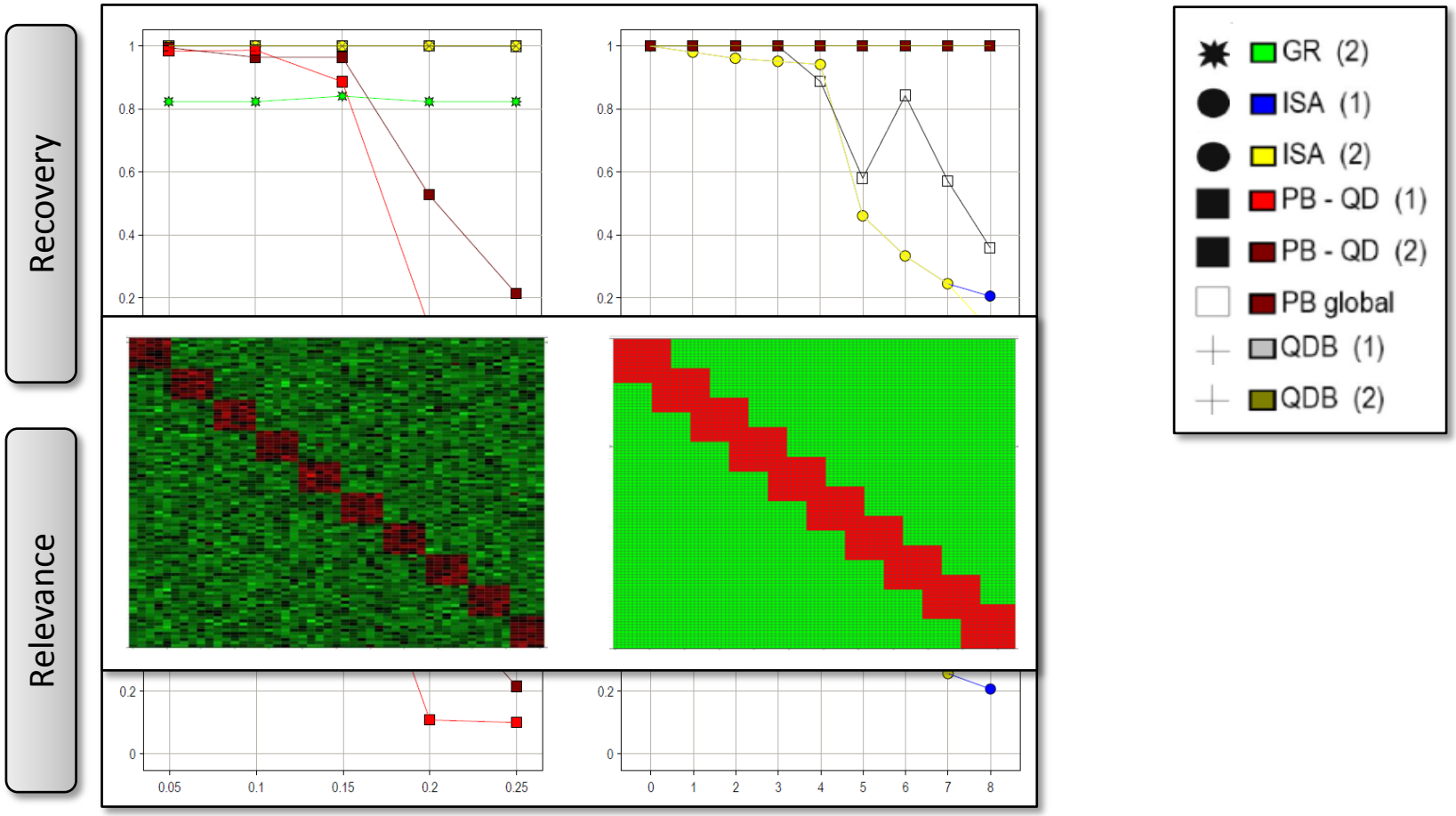
precision

recall

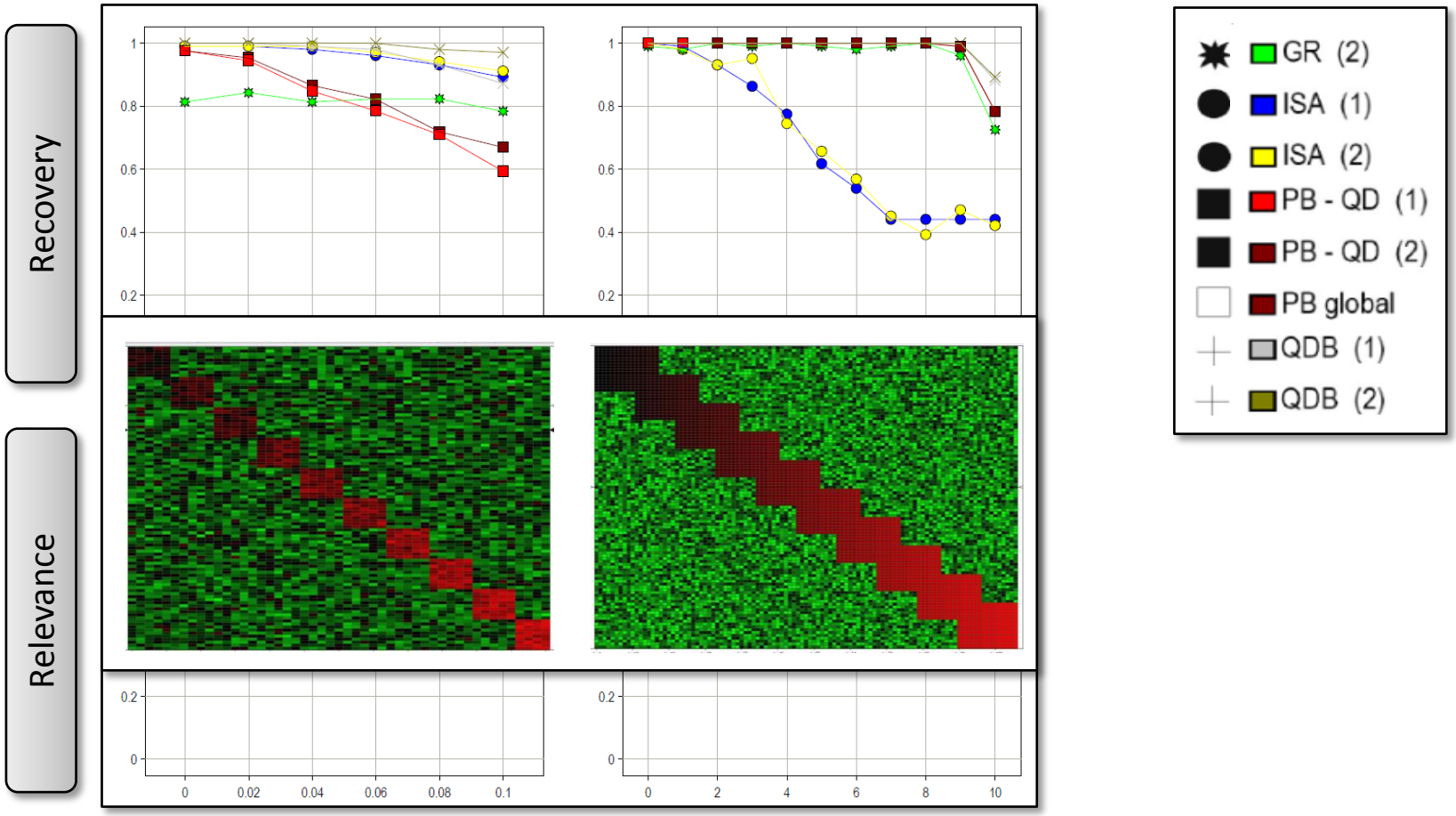
Comparison with state of the art

- Simulated datasets defined by Prelic et al. (2006)
  - Independent datasets
  - Not biased towards particular algorithm
- Query-driven biclustering algorithms:
  - Gene Recommender (GR) [Owen et al., 2003]
  - Iterative Signature Algorithm (ISA) [Ihmels et al., 2002]
  - Query-driven biclustering (QDB) [Dhollander et al., 2008]
  - *Pro*Bic (PB) [Van den Bulcke et al., 2009]

Comparison to state of the art



Comparison to state of the art

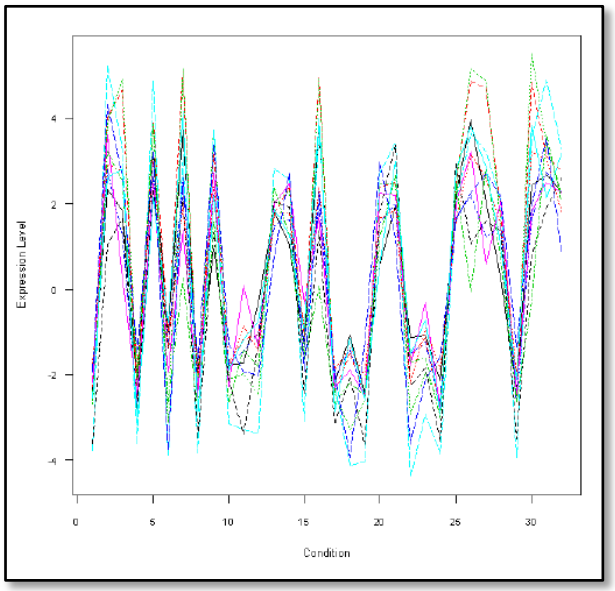


Scenario 2: constant columns

Single gene query-driven biclustering

- Select single target from a known regulator
- Use this gene as 'query' gene
- *E. coli* compendium [Lemmens et al., 2008]

Reg.	Query	Genes	Arrays	Reg. enrich.	GO term
LexA	*uvrB*	11	32	*LexA* (1.0394e-14)	SOS response (1.13e-19)
LexA	*dinI*	8	20	*LexA* (9.8312e-06)	SOS response (2.05e-12)
Fur	*fhuE*	20	75	*Fur* (1.3682e-23)	enterochelin (enterobactin) (1.43e-12)
CysB	*cysK*	12	97	*CysB* (2.3838e-20)	Sulfur metabolism (5.36e-12)
CysB	*cysD*	10	110	*CysB* (1.0744e-18)	unknown
NtrC	*ddpX*	7	28	*NtrC* (7.0782e-10)	nitrogen metabolism (5.04e-02)



Query: *uvrB*

Outlier removal in QD biclustering

Experimental setup:

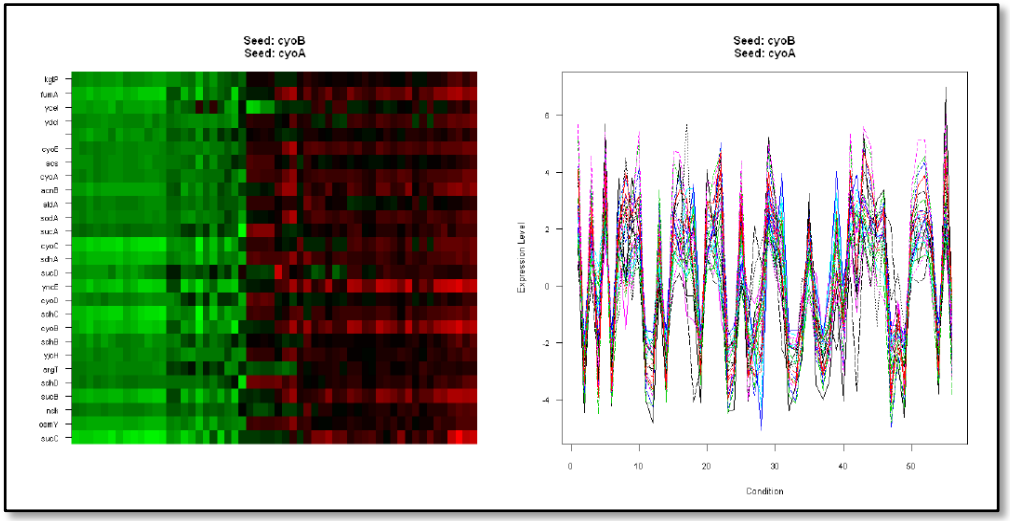
- Examine effect of outlier genes in set of query genes

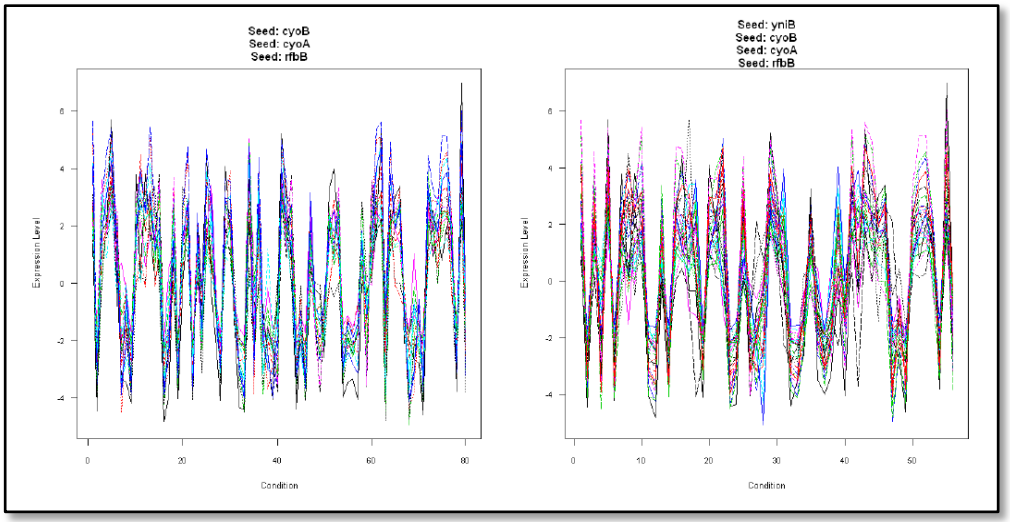
Query genes:

- Two operon genes
- One or more random genes

Example:

- *E. coli* compendium, *cyoABCDE* operon
- Select two genes (*cyoA* and *cyoB*) and one or more random genes.





Conclusion

Summary

- **SynTReN:**
  - Fast, large scale simulator of regulatory networks
  - Simulated data reveals operational characteristics of network inference algorithms unlikely to be discovered with biological data only
  - Network topology has an important effect on inference quality
- ***Pro*Bic:**
  - Combined global and query-driven biclustering model
  - Simultaneous biclustering of multiple overlapping biclusters
  - Extensibility towards module network inference
  - Robust w.r.t. noise and missing values
  - Query-driven biclustering with:
    - single genes
    - multi-gene queries containing outlier genes

Outlook & future work

- **SynTReN:**
  - Extension towards heterogeneous networks (metabolites, proteins, DNA, RNA)
- ***Pro*Bic:**
  - Development of a GUI and large scale application to biological data
  - Extend *Pro*Bic model towards regulatory module identification, including:
    - Condition annotation
    - Motif and regulator annotation

Acknowledgements

- Promotors
  - Prof. dr. ir. Bart De Moor
  - Prof. dr. ir. Kathleen Marchal
- ESAT-SCD
  Bioinformatics group
- CMPG BioI
  - ir. Hui Zhao
- ISLab
  - dr. Koen Van Leemput

Publications

Published

- T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor and K. Marchal. *SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms.* BMC Bioinformatics 2006, 7:43.
- T. Van den Bulcke, K. Lemmens, Y. Van de Peer, and K. Marchal. *Inferring transcriptional networks by mining 'omics' data.* Current Bioinformatics 2006, 1:3.
- T. Michoel, S. Maere, E. Bonnet, A. Joshi, T. Van den Bulcke, K. Van Leemput, P. van Remortel, M. Kuiper, K. Marchal and Y. Van de Peer. *Validating module network learning algorithms using simulated data.* BMC Bioinformatics 2007, 8(Suppl 2):S5.
- K. Van Leemput, T. Van den Bulcke, T. Dhollander, B. De Moor, K. Marchal, P. van Remortel. *Exploring the operational characteristics of inference algorithms for transcriptional networks by means of synthetic data.* Artificial Life 2008, 14:1, 49-63.

Submitted

- H. Sung, K. Lemmens, T. Van den Bulcke, K. Engelen, B. De Moor, K. Marchal. *ViTraM: Visualization of Transcriptional Modules*, Bioinformatics 2009.

In preparation

- T. Van den Bulcke, H. Zhao, K. Engelen, B. De Moor and K. Marchal. *Pro*Bic*: Global and query-driven biclustering of gene expression data using Probabilistic Relational Models.*