From genomes to networks by using tree-based supervised learning methods

GENIE3 within the DREAM4 and DREAM5 challenges

Pierre Geurts

Dept. of EECS & GIGA-R, University of Liège, Belgium

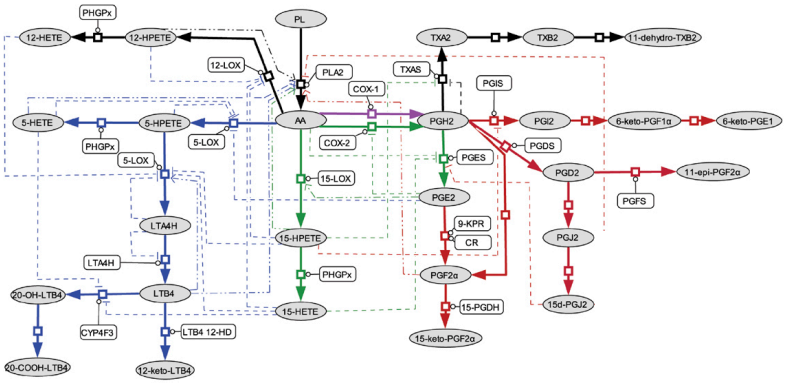
BioMAGNet Annual Meeting 2011
Monday, March 21st, 2011

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel (ULg)
Yvan Saeys (PSB, UGent)
Jimmy Vandel (INRA Toulouse)

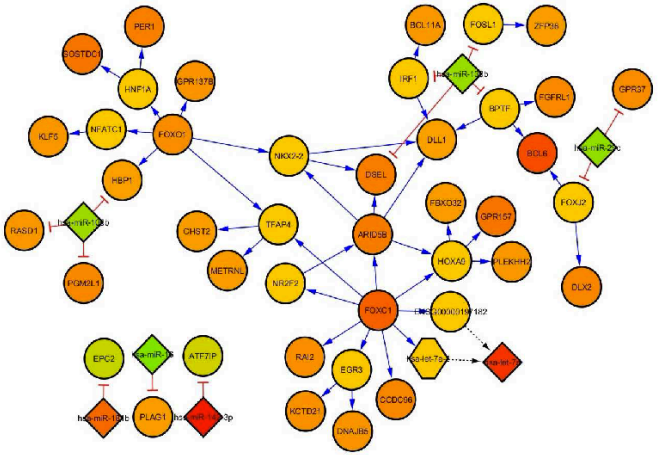
Biological networks

Networks or graphs are very common to represent biological information

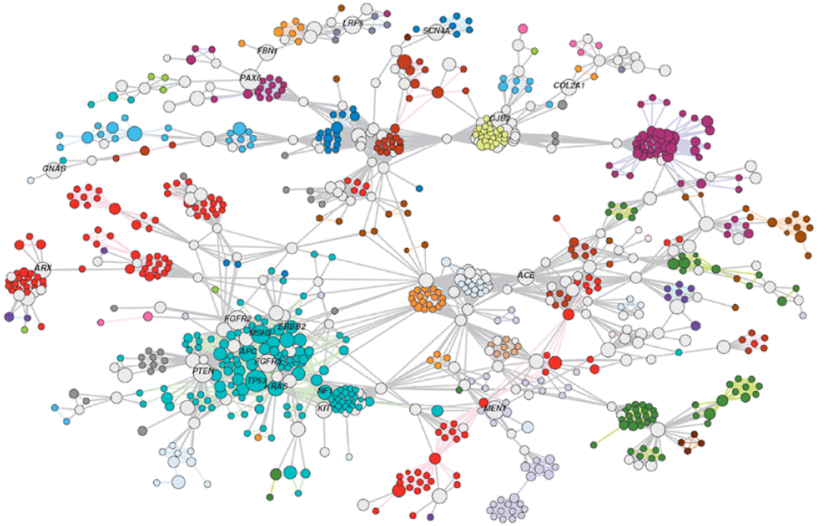
Metabolic network



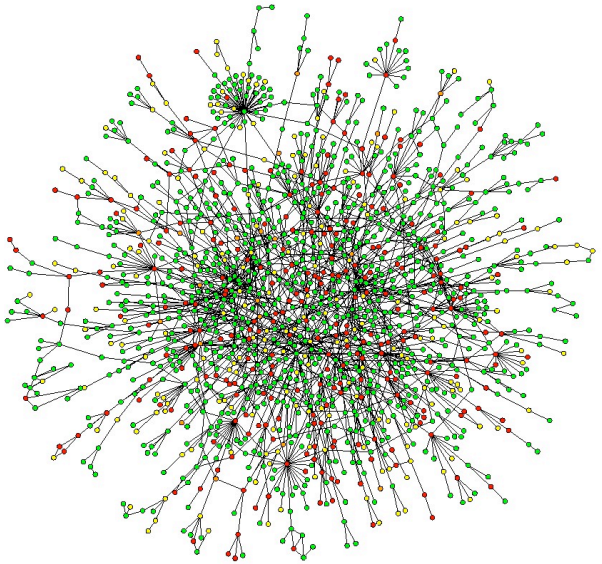
Gene regulatory network



Disease gene network



Protein-protein interaction network



Biological network inference

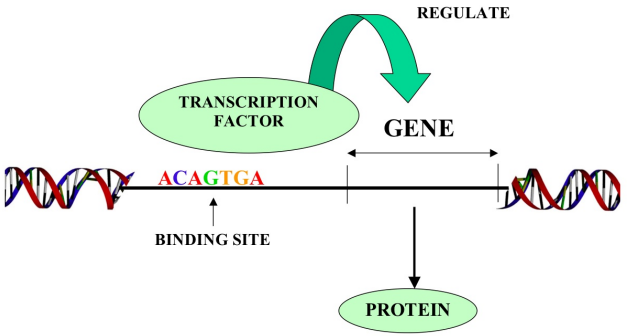
- The knowledge about these networks is typically obtained through wet-lab (small or large-scale) experiments

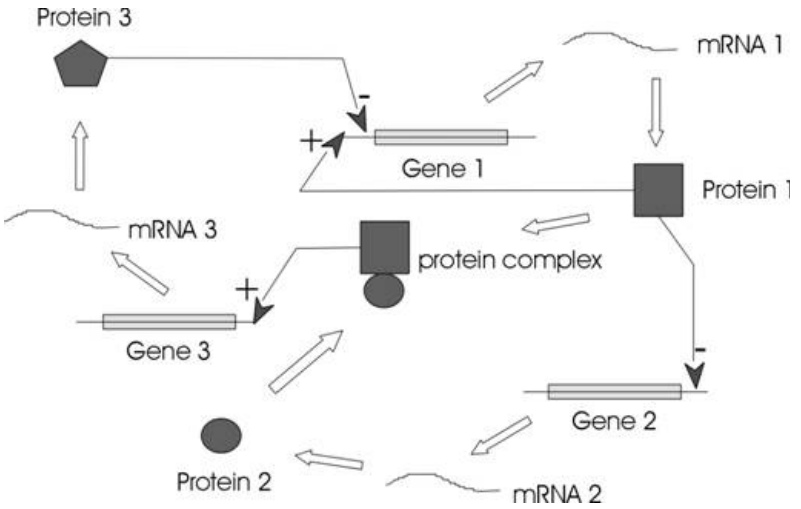
$\Rightarrow$ partial, noisy, costly

- Experimental techniques are usefully complemented by computational inference methods
- Motivation:
  - Predict novel interactions
  - Confirm/invalidate experimental predictions
  - Explain known interactions from different points of view
  - Inference of properties of "new" genes/proteins

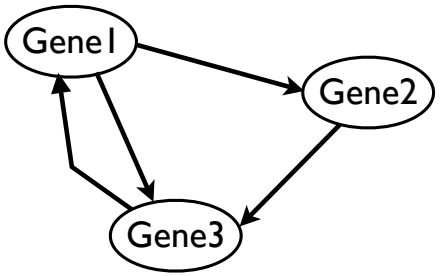
Regulatory network inference

Transcription regulation

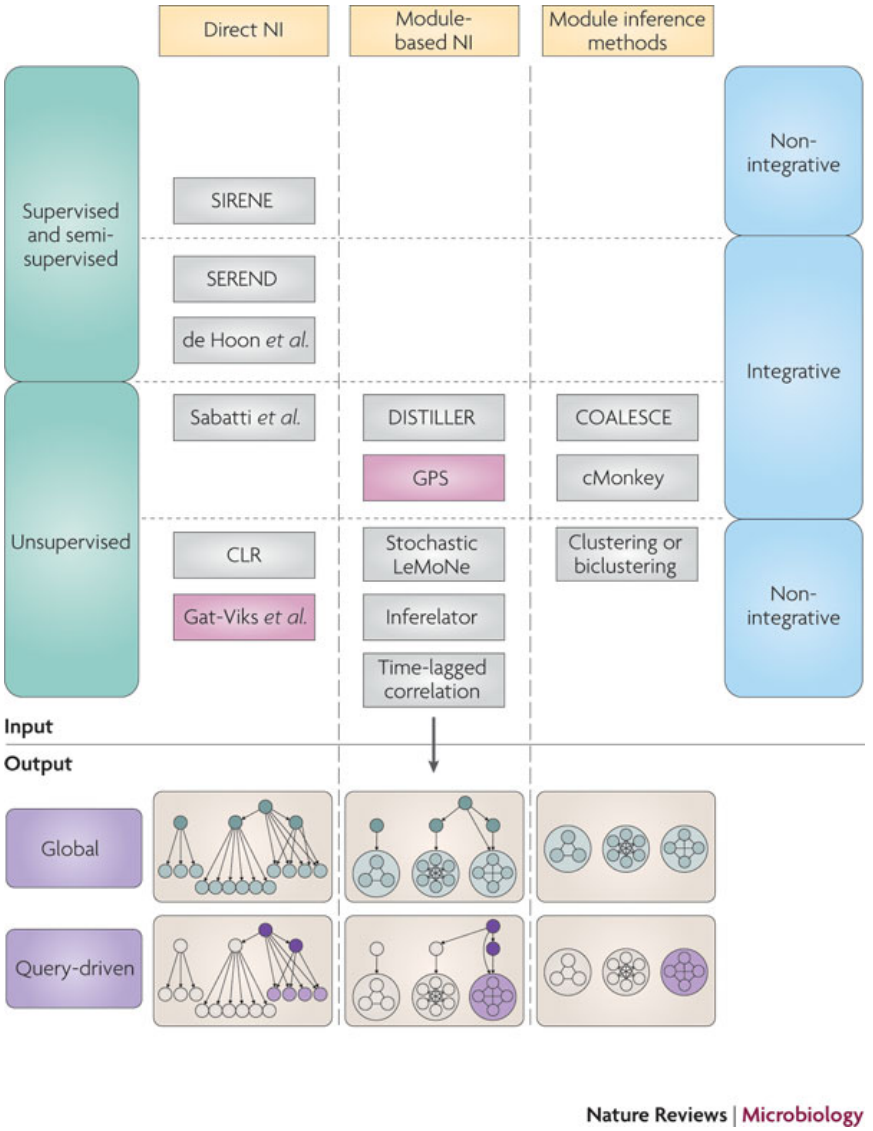




Simplified view for inference

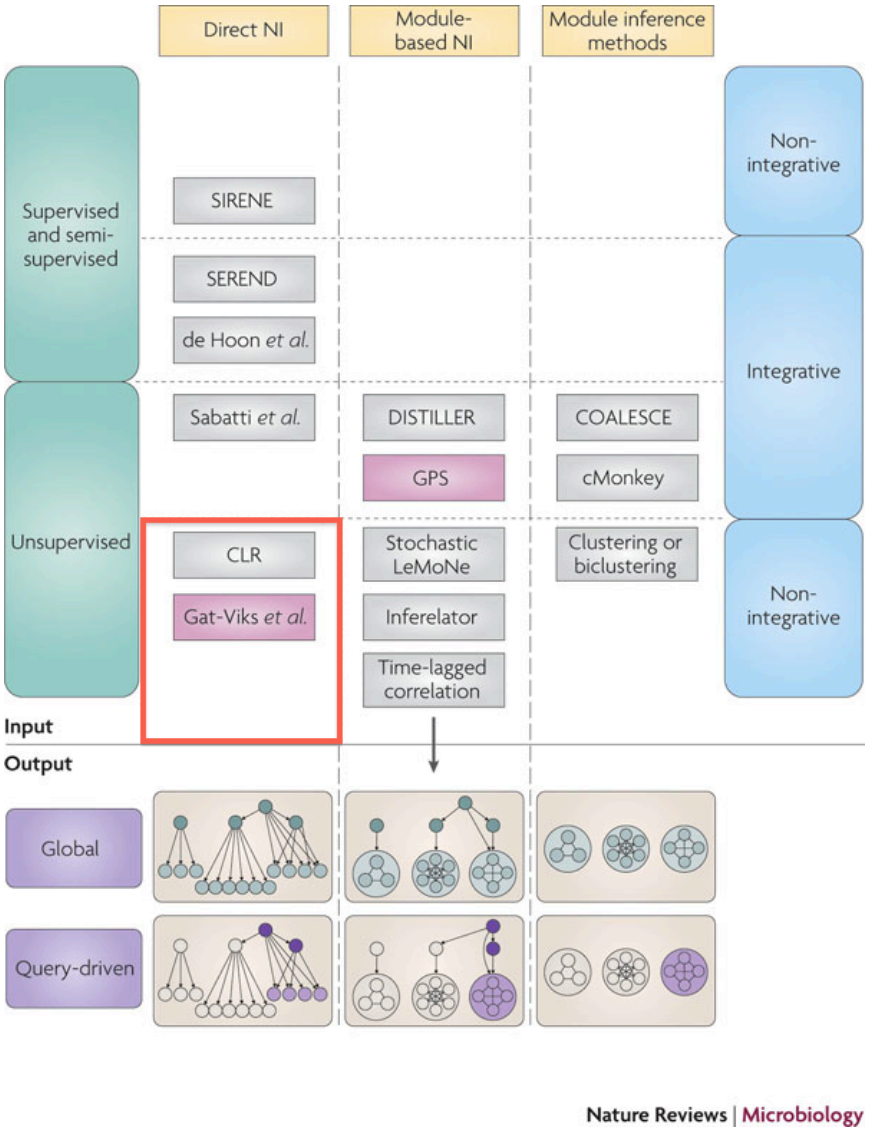


Regulatory network inference methods



(De Smet & Marchal, Nature Review Microbiology, 2010)

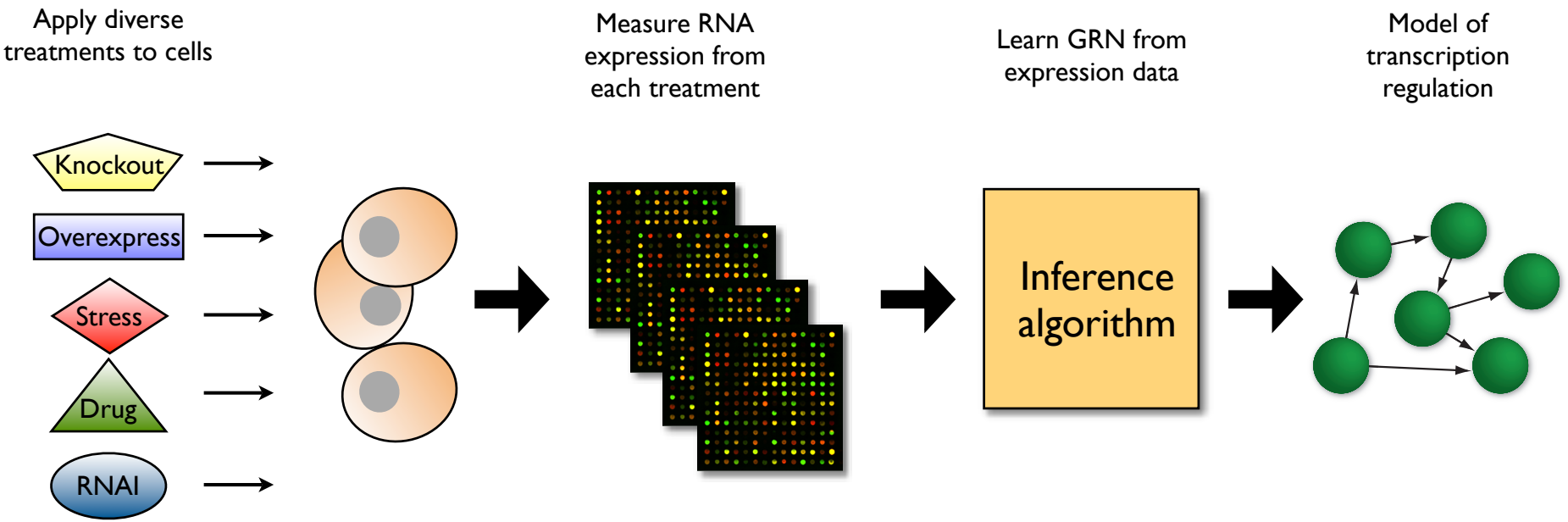
Regulatory network inference methods



(De Smet & Marchal, Nature Review Microbiology, 2010)

Outline

Gene regulatory network inference from expression data



GRN inference: issues

Several issues for GRN inference algorithms:

- Very heterogeneous data, far from i.i.d. (knock-out, time series, multifactorial, etc.)
- Directing the edges is difficult (esp. with static data)
- Direct vs indirect interactions
- Large p/small n problem
- Scalability (thousands of genes, millions of pairs)
- Edge ranking vs network prediction
- Difficult to validate (very few known real networks)

GRN inference: existing methods

Two main families of methods:

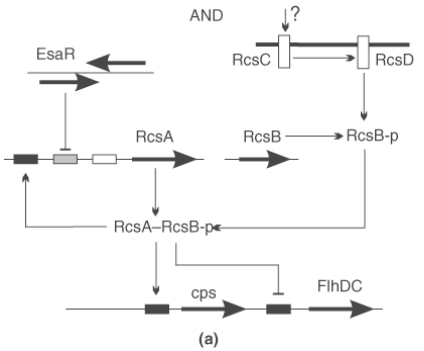
- Score-based: define a similarity score between genes based on their expression profiles and connect two genes if their similarity is above some threshold

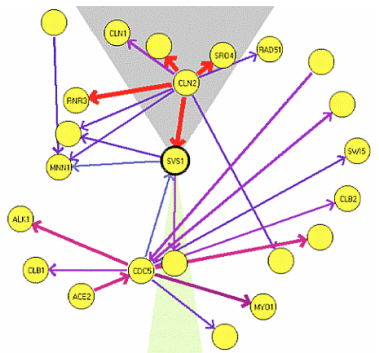
Score matrix

Regulating gene \ Target gene	gene 1	gene 2	· · ·	gene p
gene 1	-	0.05	· · ·	0.56
gene 2	0.19	-	· · ·	0.03
· · ·	· · ·	· · ·	· · ·	· · ·
gene p	0.11	0.42	· · ·	-

(eg., CLR, MRNET, ARACNE, GeneNet)

- Model-based: learn a model that explains as well as possible the observed expression data and extract the network from this model



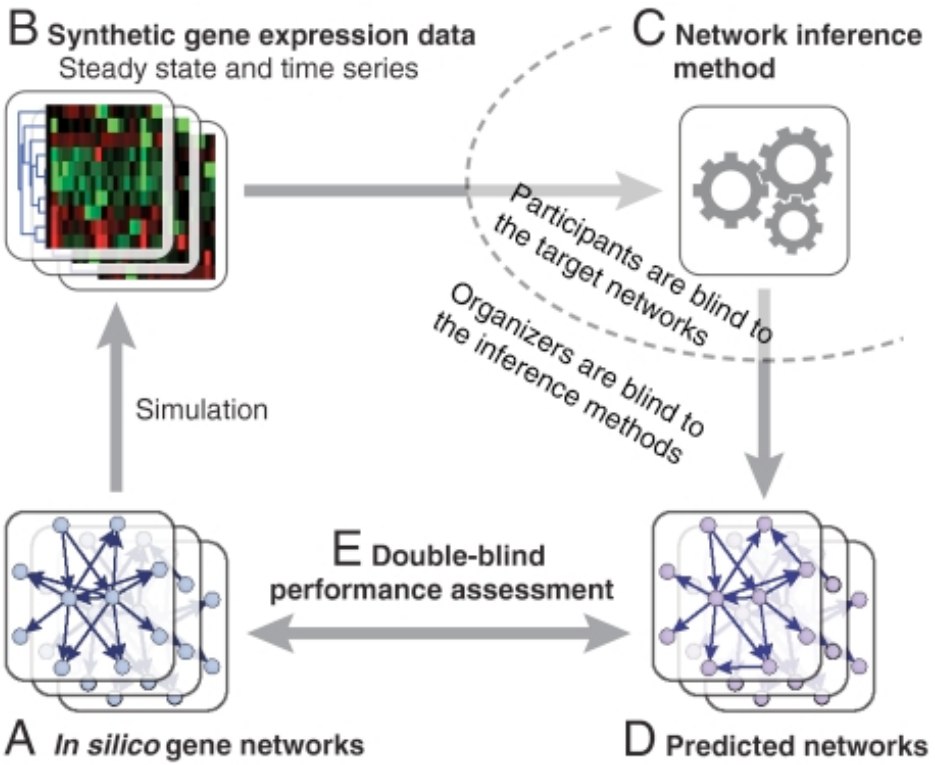


$$\frac{dx_i}{dt} = m_i \cdot f_i(\mathbf{y}) - \lambda_i^{\text{RNA}} \cdot x_i$$

$$\frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{\text{Prot}} \cdot y_i,$$

(eg., Boolean networks, differential equations, (dynamic) Bayesian networks)

DREAM challenges

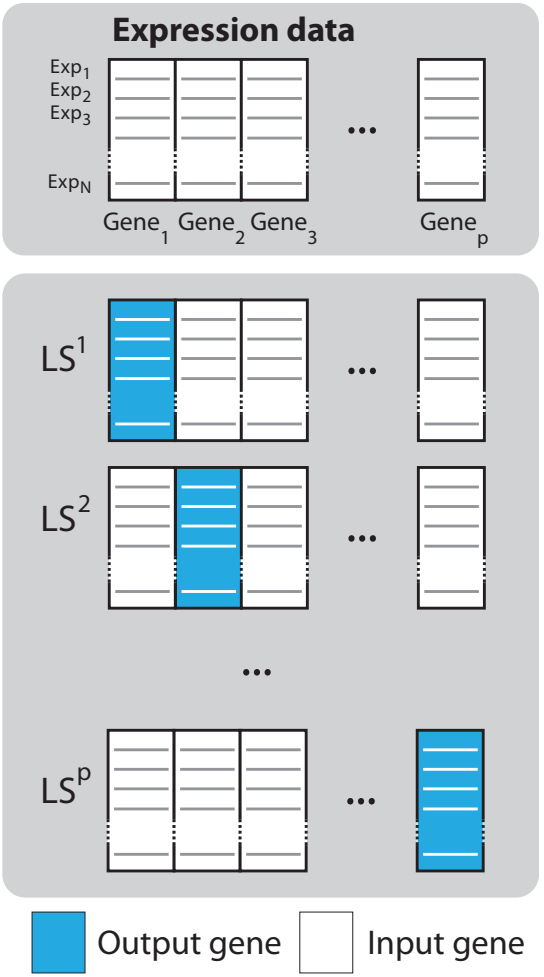


(Marbach et al., PNAS, 2010)

- Very few reliable real benchmark datasets exist $\Rightarrow$ evaluation on simulated data
- DREAM, "Dialogue for Reverse Engineering Assessments and Methods", is an annual reverse engineering competition, organized since 5 years (`http://wiki.c2b2.columbia.edu/dream`)

Outline

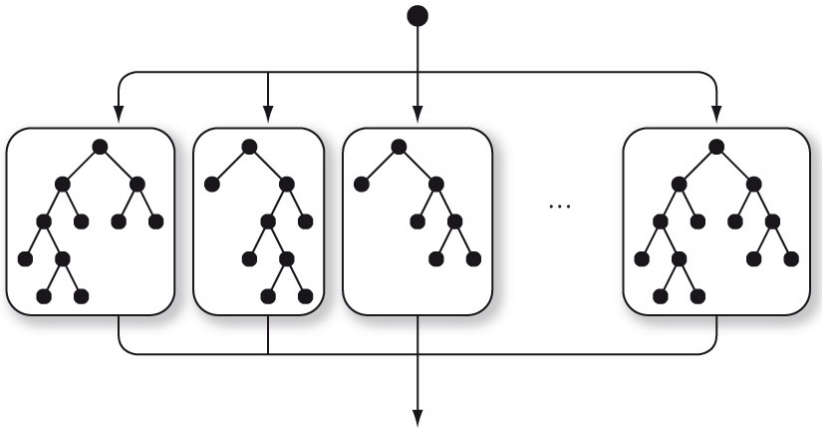
Network inference as p feature selection problems



- Main idea: decompose the problem of network inference into p sub-problems
- Sub-problem i = find the regulators of gene i, ie., those genes whose expression is predictive of gene i's expression
- Solved as p feature selection problems (in regression)

Feature selection with tree-based ensemble methods

Tree-based ensemble methods are good candidates



Bagging
Random Forests
Extra-Trees
...

Non-parametric models

Can deal with interacting features

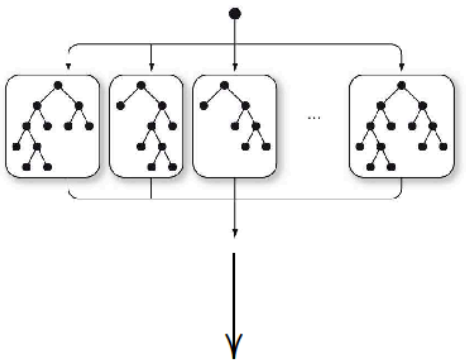
(Almost) parameter-free

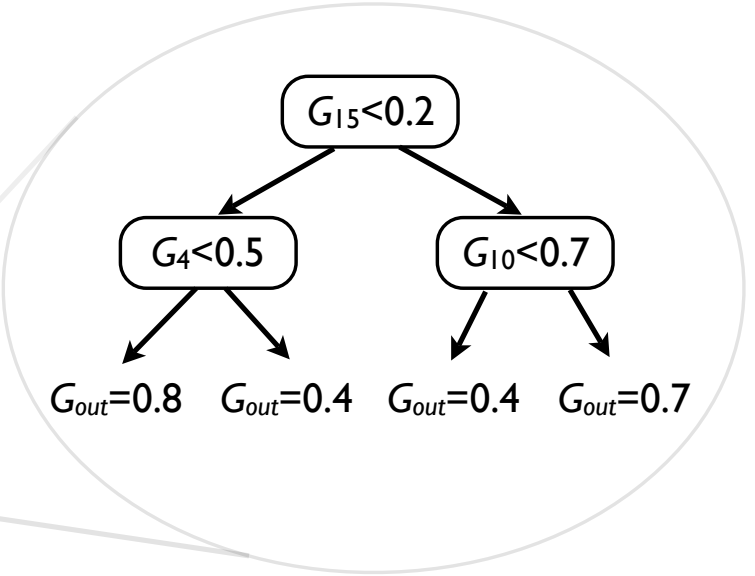
Work well with high-dimensional datasets

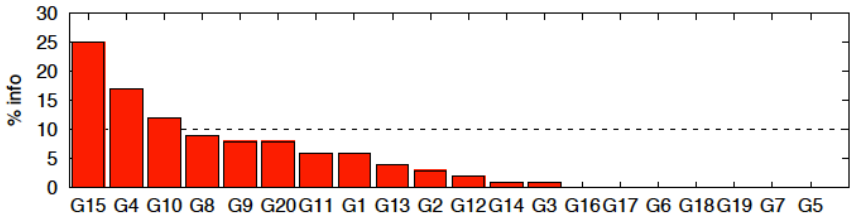
Scalable

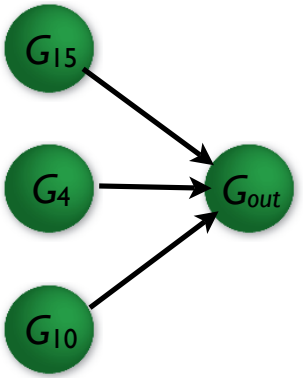
Our approach

G_1	G_2	G_3	$\cdots$	G_n	G_{out}
0.106	0.878	0.054	$\cdots$	0.870	0.899
0.014	0.860	0.031	$\cdots$	0.890	0.919
0.062	0.443	0.158	$\cdots$	0.877	0.957
0.011	0.896	0.002	$\cdots$	0.882	0.945
0.076	0.783	0.000	$\cdots$	0.883	0.932
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
0.079	0.892	0.005	$\cdots$	0.862	0.912



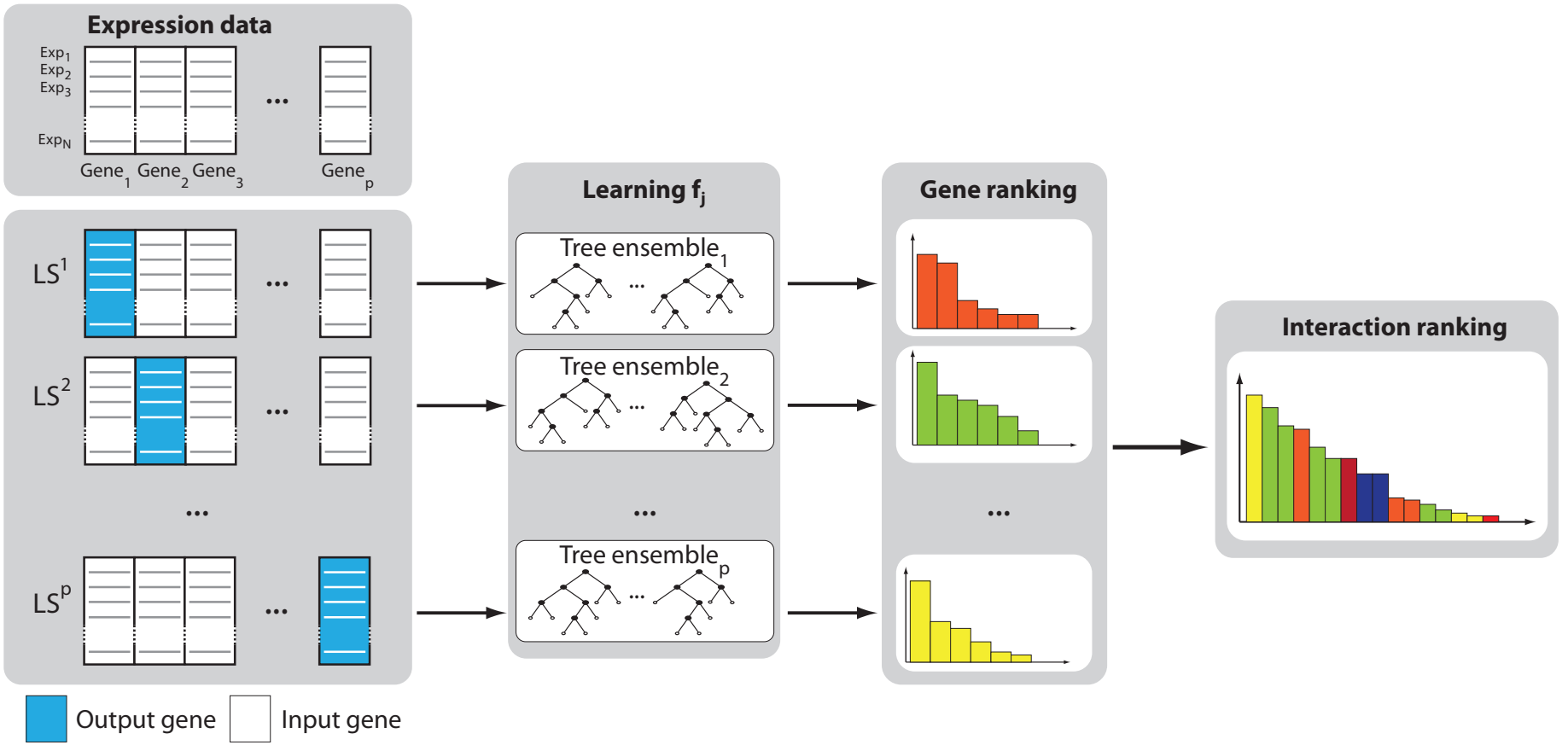






(Bagging, with 1000 trees in all experiments)

GENIE3 (GEne Network Inference with Ensemble of Trees)



Discussion

GENIE3:

- Extends score-based methods by taking into account variable dependencies
- Can be considered as a non-parametric model-based approach, related to Bayesian networks
- Has a reasonable computational complexity (at most $O(p^2 N \log(N))$) and is trivially parallelizable

Outline

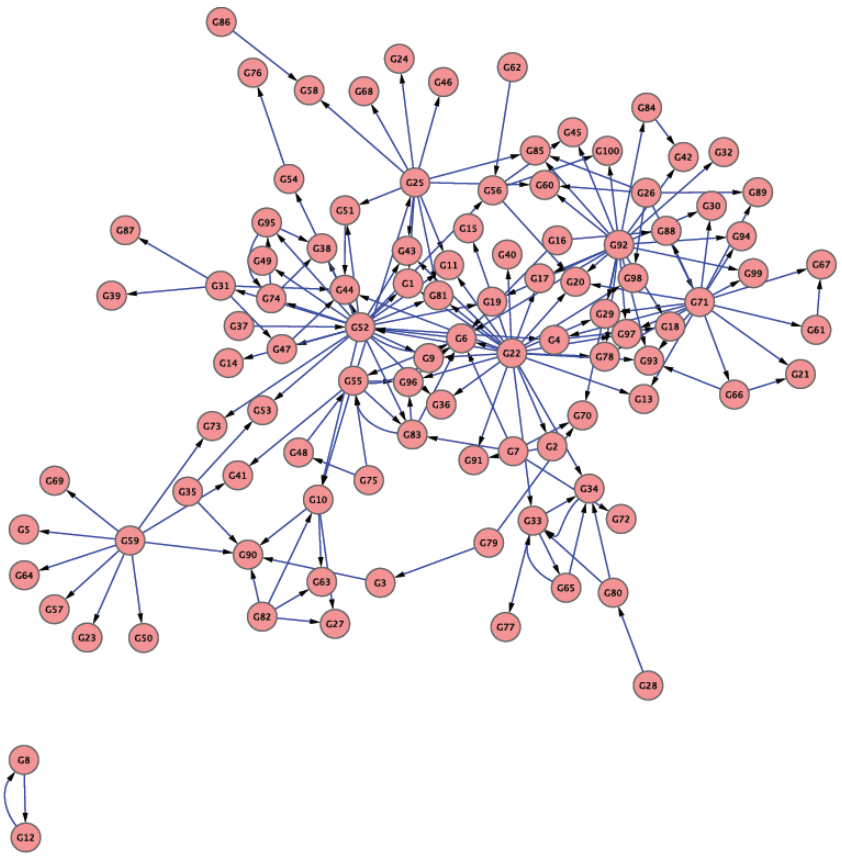
Experiments within the DREAM challenges

- DREAM3 (2008):
  - In-Silico-Network challenge (Size 100): 5 networks of 100 genes, data: time series+knock-down+knock-out
- DREAM4 (2009):
  - In Silico Size 100: 5 networks of 100 genes, data: time series+knock-down+knock-out
  - In Silico Size 100 Multifactorial: 5 networks of 100 genes, steady-state levels under multifactorial pertubations
- DREAM5 (2010):
  - Network Inference challenge: 3 real networks + 1 artificial network, data: microarray compendia
  - Systems Genetics challenge: 5×3 networks of 100 genes, data: gene expression and genotyping data

Steady-state data (and microarray compendia)

- DREAM3 (2008):
  - In-Silico-Network challenge (Size 100): 5 networks of 100 genes, data: time series+knock-down+knock-out
- DREAM4 (2009):
  - In Silico Size 100: 5 networks of size 100, data: time series+knock-down+knock-out
  - In Silico Size 100 Multifactorial: 5 networks of size 100, steady-state levels under multifactorial pertubations
- DREAM5 (2010):
  - Network Inference challenge: 3 real networks + 1 artificial network, data: microarray compendia
  - Systems Genetics challenge: 5×3 networks of 100 genes, data: gene expression and genotyping data

DREAM4 *In silico multifactorial* challenge

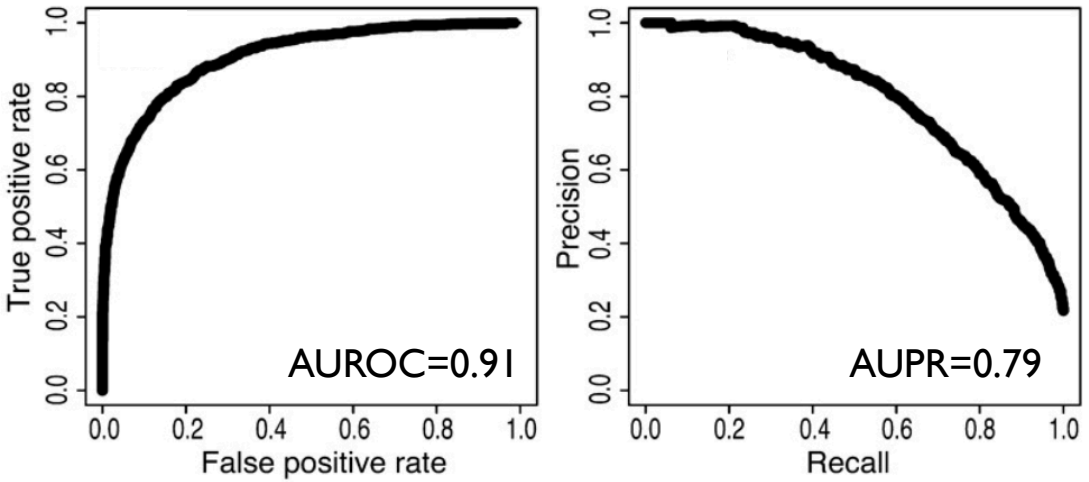


- 5 networks of 100 genes each, extracted from real GRN of *E. coli* and *S. cerevisiae*
- Detailed kinetic model in the form of (stochastic) ordinary differential equations (plus noise)

$$\frac{dx_i}{dt} = m_i f_i(y) - \lambda_i^{\text{RNA}} x_i$$
$$\frac{dy_i}{dt} = r_i x_i - \lambda_i^{\text{Prot}} y_i$$

- 100 expression measurements: static steady-state expression profiles obtained from (slight) perturbations of the basal activation of all genes

Evaluation protocol (all challenges)



- Output of algorithms: a ranked list of predicted interactions (directed)
- Evaluation through ROC and Precision-recall curves

$\Rightarrow$ Area under ROC (AUROC) and Precision-recall (AUPR) curves

$\Rightarrow$ p-values under random model

- Overall score $= -0.5 \log_{10}(p_{roc} p_{pr})$

Our Results on DREAM4

Final ranking of the challenge (*directed* network)

Rank	Team	Overall Score	Mean AUPR	p-value	Mean AUROC	p-value
1	GENIE3-Bagging	37.736	0.22	5.93e-54	0.76	1.93e-28
2	Team 549	28.165	0.14	7.45e-35	0.73	6.29e-23
...	...	...	...	...	...	

Comparison with existing approaches (*undirected* network)

	GENIE3-Bagging	CLR	ARACNE	MRNET	GGM
Overall score	36.736	35.838	32.632	34.124	26.846

GENIE3 is able to predict a *directed* network

Predicted networks contain a significant number (52%) of asymmetric links (versus 95% in the gold standard).

At 5% (resp. 100%) recall, mean error rate on edge directionality is 20% (resp. 25%)
(edges $i \rightarrow j$ for which $w_{i \rightarrow j} < w_{j \rightarrow i}$).

$\rightarrow$ GENIE3 is a plausible approach for directing an undirected network.

DREAM5 *Network Inference* Challenge

DREAM4 datasets are not realistic

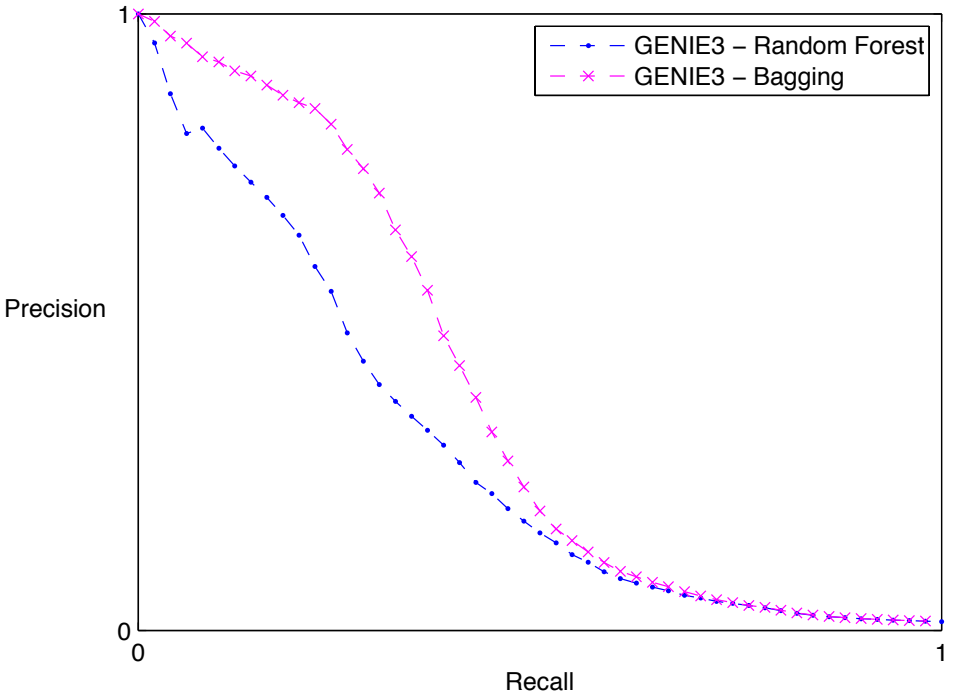
- i.i.d. multifactorial data
- Small number of genes
- Number of genes $\simeq$ number of experiments

DREAM5 data:

- 3 real networks: *E. coli*, *S. cerevisae*, *S. aureus* (no gold standard for the last one but community predictions will be verified experimentally).
- 1 simulated network: same simulation model as in DREAM4 but mimic main features of real microarray compendia
- **Potential TFs are supposed to be known in advance**

Network	# TFs	# Genes	# Chips
in-silico	195	1643	805
S. aurus	99	2810	160
E. coli	334	4511	805
S. cerevisiae	333	5950	536

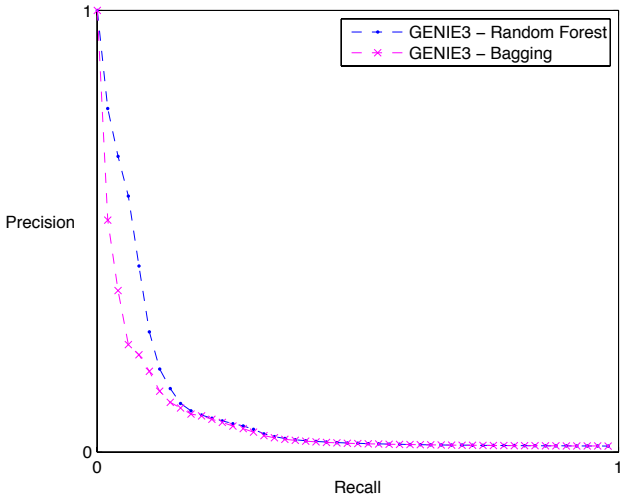
DREAM5 - *In silico* network



Team	AUPR	AUROC
GENIE3-Bag	**0.38**	**0.82**
Team 862	0.31	0.76
Team 776	0.30	0.78
GENIE3-RF	**0.29**	**0.82**
Team 868	0.28	0.74
Team 870	0.28	0.75
...	...	...

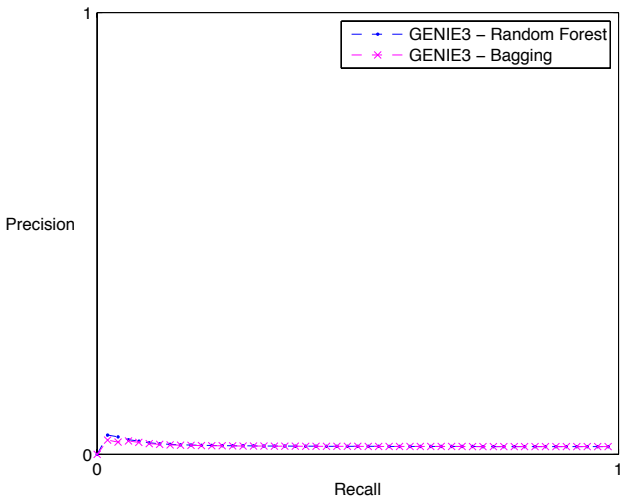
DREAM5 - *In vivo* networks

E. coli



Team	AUPR	AUROC
Team 543	0.12	0.67
GENIE3-RF	**0.09**	**0.62**
Team 772	0.09	0.61
Team 48	0.09	0.61
Team 395	0.08	0.60
...	...	...
GENIE3-Bag	**0.07**	**0.61**
...	...	...

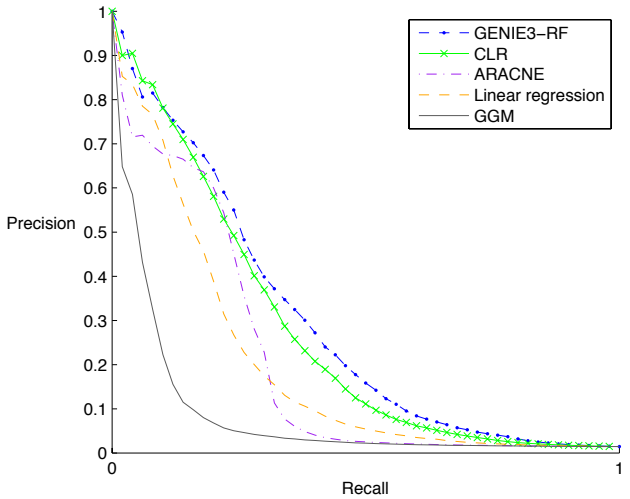
S. Cerevisiae



Team	AUPR	AUROC
Team 702	0.03	0.51
Team 548	0.03	0.51
Team 395	0.03	0.54
Team 705	0.03	0.52
...	...	...
GENIE3-RF	**0.02**	**0.52**
GENIE3-Bag	**0.02**	**0.52**
...	...	...

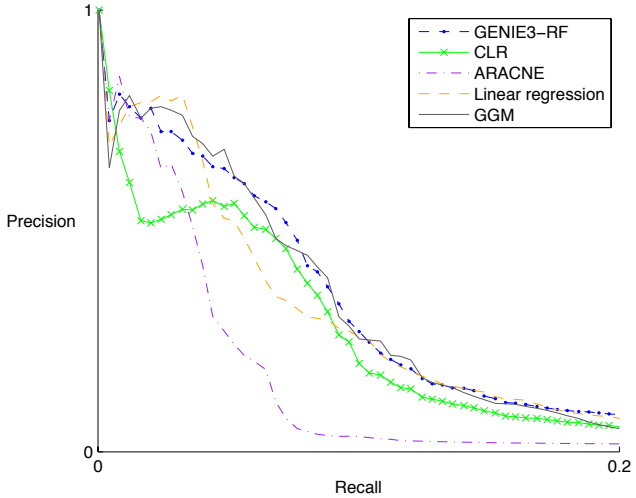
Comparison with other methods

In silico

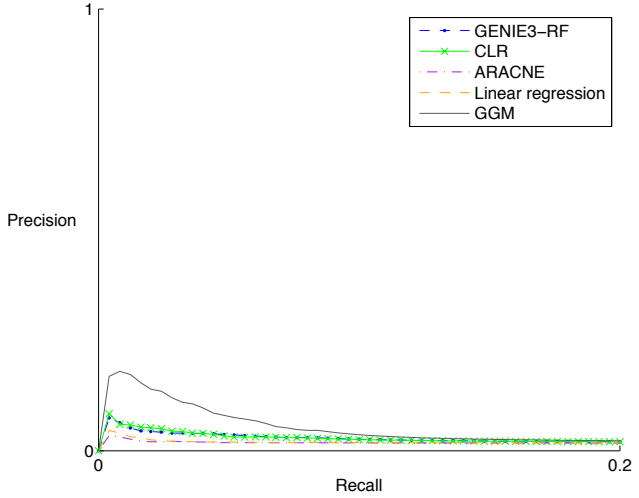


	Overall Score
GENIE3-RF	40.28
2nd DREAM5	34.02
CLR	23.93
Linear reg.	7.15
GGM	5.81
ARACNE	3.22

E. coli



S. Cerevisiae



CLR: Faith et al. (2007) ARACNE: Margolin et al. (2006) GGM: Schafer et al. (2005)

Time-series (and steady-state) data

- DREAM3 (2008):
  - In-Silico-Network challenge (Size 100): 5 networks of 100 genes, data: time series+knock-down+knock-out
- DREAM4 (2009):
  - In Silico Size 100: 5 networks of size 100, data: time series+knock-down+knock-out
  - In Silico Size 100 Multifactorial: 5 networks of size 100, steady-state levels under multifactorial pertubations
- DREAM5 (2010):
  - Network Inference challenge: 3 real networks + 1 artificial network, data: microarray compendia
  - Systems Genetics challenge: 5×3 networks of 100 genes, data: gene expression and genotyping data

GENIE3 with time-series data

Time-series only:

- Predict expressions at time t from previous time steps, i.e., minimizes for each gene j:

$$\sum_t (g_j(t+h) - f_j(g^{-j}(t)))^2$$

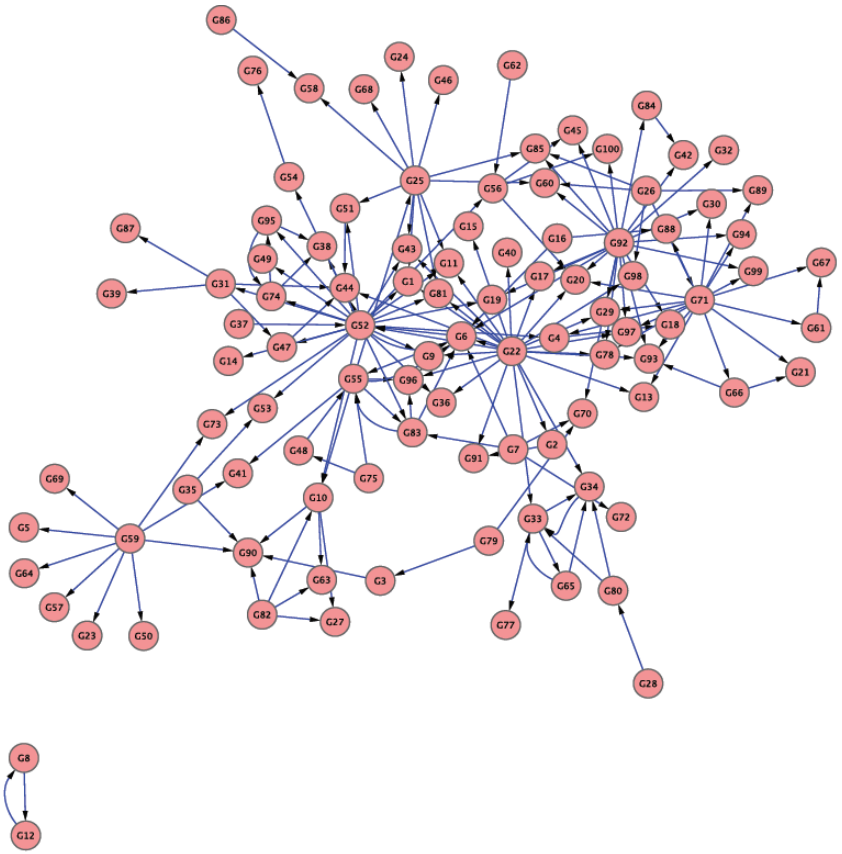
$g_1(t_1)$	$g_2(t_1)$	$\cdots$	$g_p(t_1)$	$g_{out}(t_1+h)$
$g_1(t_2)$	$g_2(t_2)$	$\cdots$	$g_p(t_2)$	$g_{out}(t_2+h)$
$g_1(t_3)$	$g_2(t_3)$	$\cdots$	$g_p(t_3)$	$g_{out}(t_3+h)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

(averaging over several time horizons h works best)

Time-series plus steady-state:

- Learn a *separate* ranking from both datasets and then combine them
- or *Jointly* learn a single model for both datasets by merely concatenating them

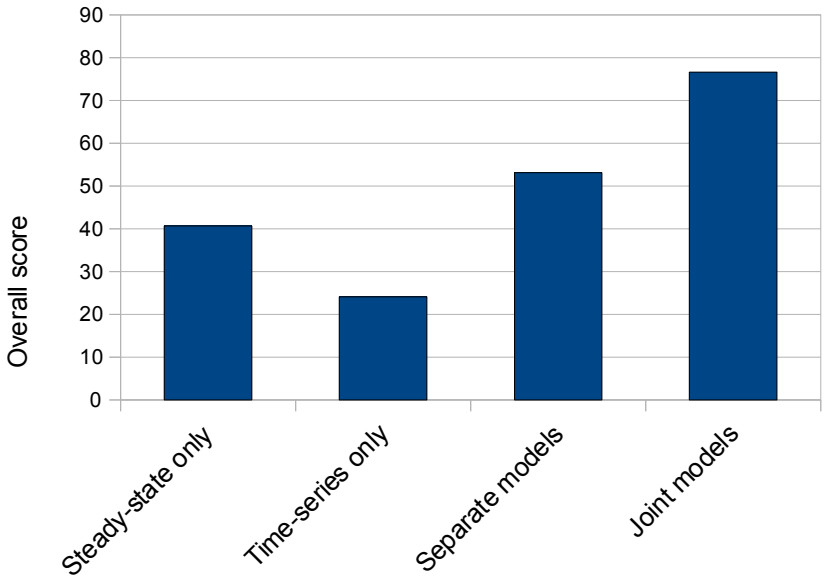
DREAM3 and DREAM4 In Silico Size100



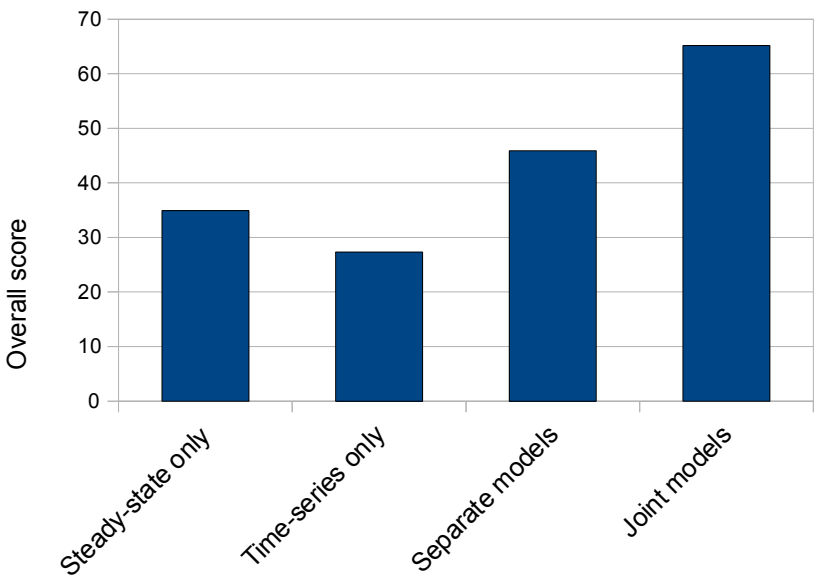
- 5 networks of 100 genes each, extracted from real GRN of *E. coli* and *S. cerevisiae*
- Steady-state data: 201 profiles (systematic knock-down and knock-out of all genes, wild-type)
- Time-series: 210 profiles (10×21 time points)

Results

DREAM3 Size 100



DREAM4 Size 100

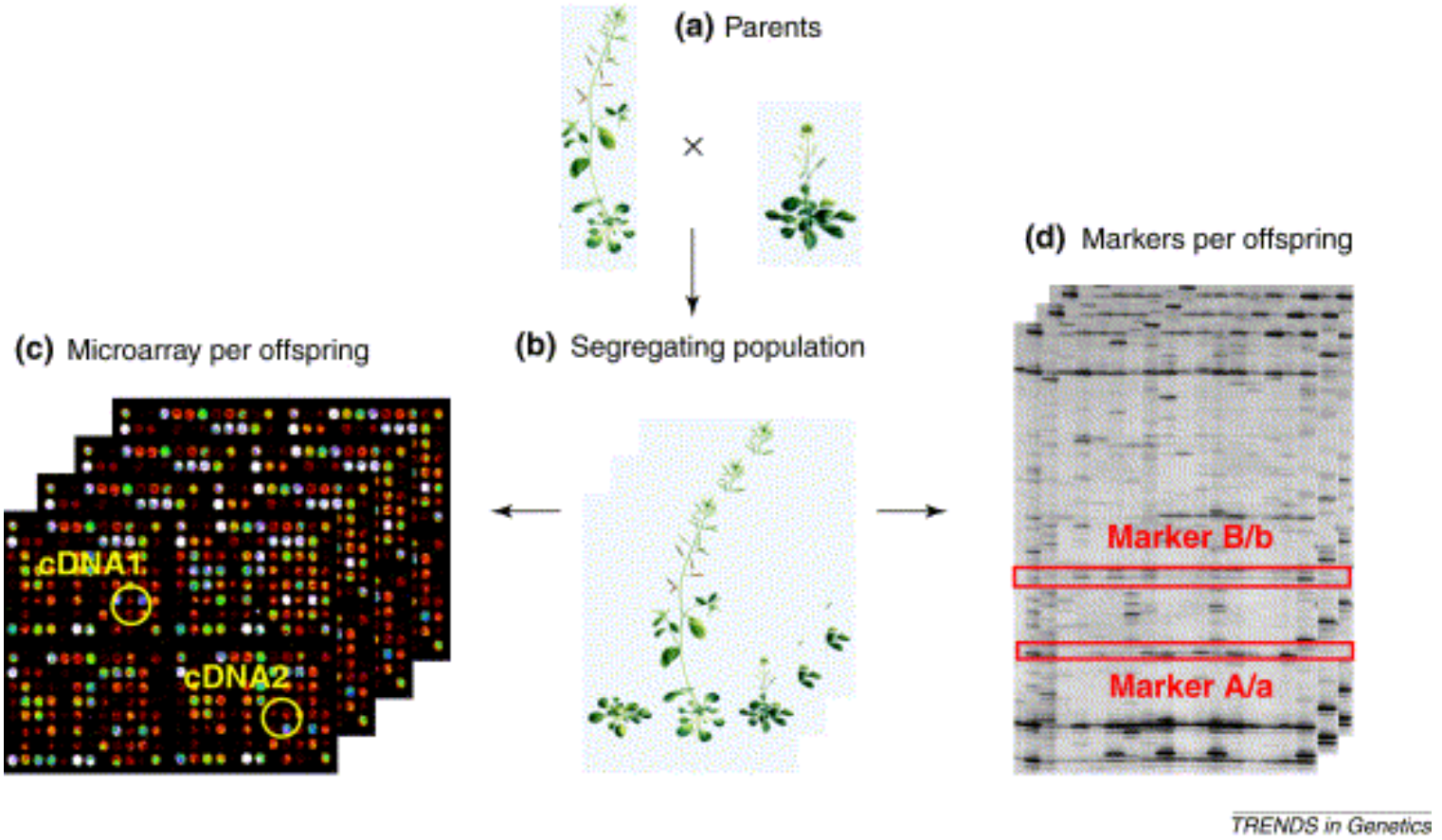


We would have been ranked 2nd on DREAM3 and 3rd on DREAM4.
(Best methods are based on dynamical models.)

Expression + Genotyping data

- DREAM3 (2008):
  - In-Silico-Network challenge (Size 100): 5×3 networks of 100 genes, data: time series+knock-down+knock-out
- DREAM4 (2009):
  - In Silico Size 100: 5 networks of size 100, data: time series+knock-down+knock-out
  - In Silico Size 100 Multifactorial: 5 networks of size 100, steady-state levels under multifactorial pertubations
- DREAM5 (2010):
  - Network Inference challenge: 3 real networks + 1 artificial network, data: microarray compendia
  - Systems Genetics challenge: 5×3 networks of 100 genes, data: gene expression and genotyping data

Systems genetics



(Jansen and Nap, Trends in Genetics, 2001)

Conclusions and future works

Conclusions:

- We obtained very good performances with GENIE3 on the DREAM challenges with different kinds of data
- Performances on real datasets are worse than expected from results on artificial data
- The availability of dynamical models (synthetic and real ones) is crucial to fairly assess and thus design network inference methods

Future works:

- Investigation of potential differences between real and artificial datasets
- Incorporate other kinds of regulations (miRNAs)
- Combination with dynamical models (in both directions)
- Application on real datasets

References

GENIE3 (steady-state, DREAM4):

V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts.

Inferring regulatory networks from expression data using tree-based methods.

PLoS ONE, 5(9):e12776, 2010.

Software:

`http://www.montefiore.ulg.ac.be/~huynh-thu/software.html`

DREAM challenges:

`http://wiki.c2b2.columbia.edu/dream`