Inferring regulatory networks from expression data using tree-based methods

Vân Anh HUYNH-THU,
Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts

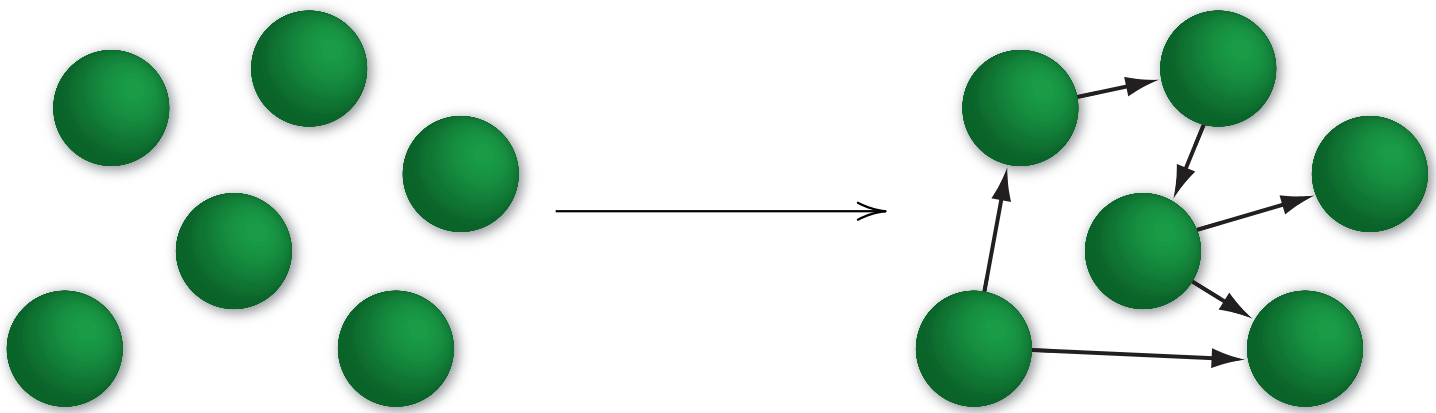
MLSB 2010, Edinburgh
October 2010

Department of Electrical Engineering and Computer Science,
Systems and Modeling
GIGA-Research, Bioinformatics and Modeling
University of Liège, Belgium

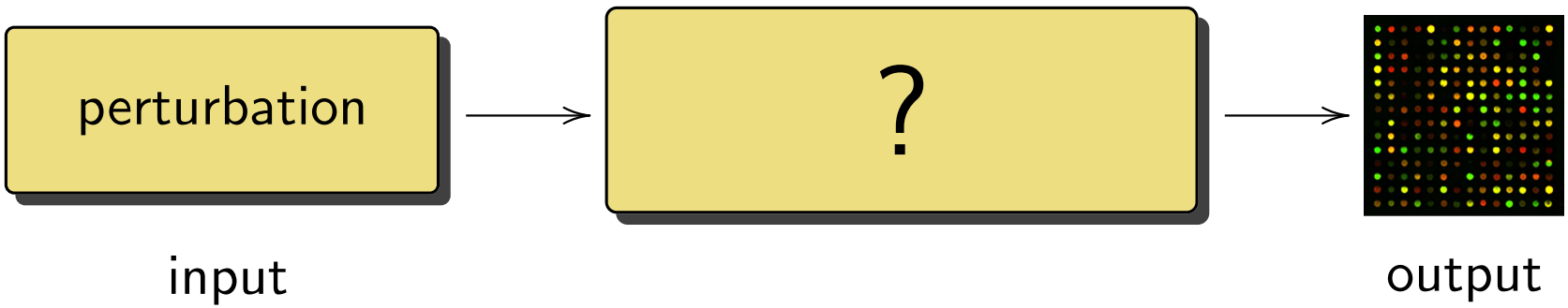
Inferring regulatory networks is a challenging problem

unknown network

inferred network

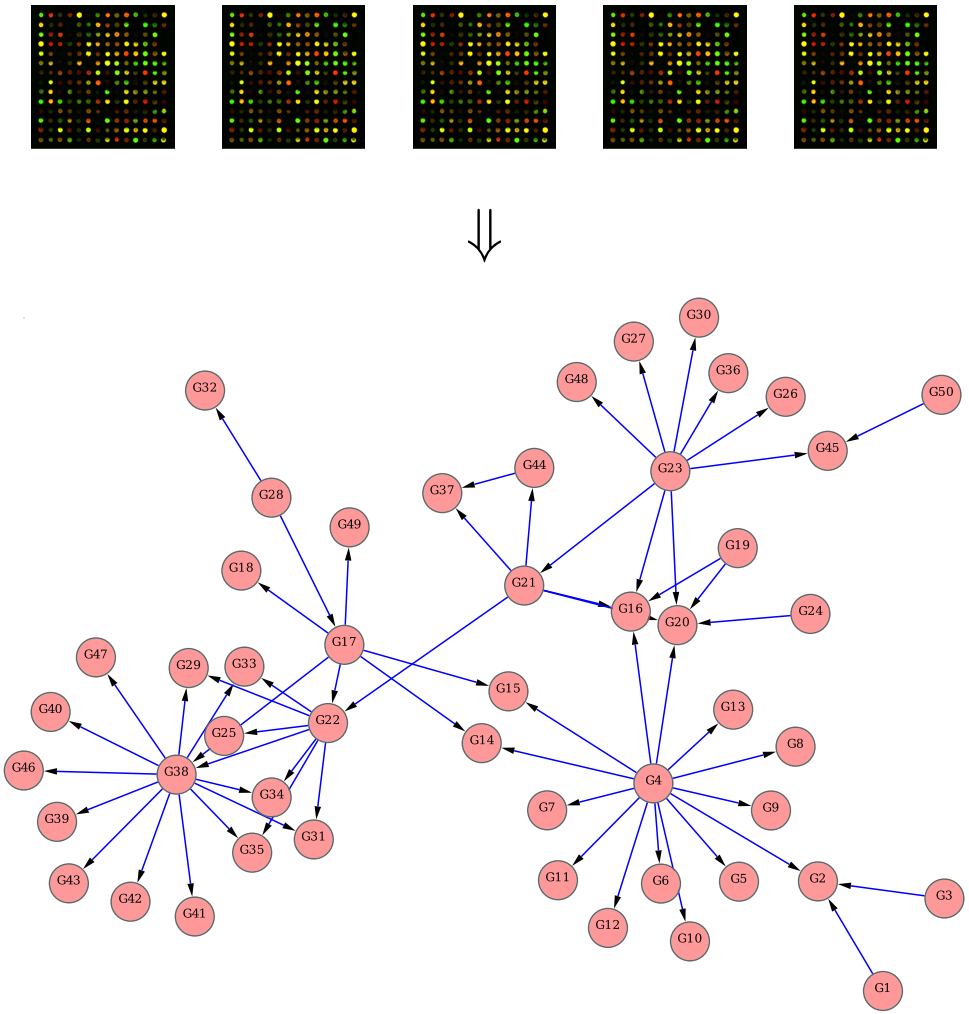


Gene expression can be measured

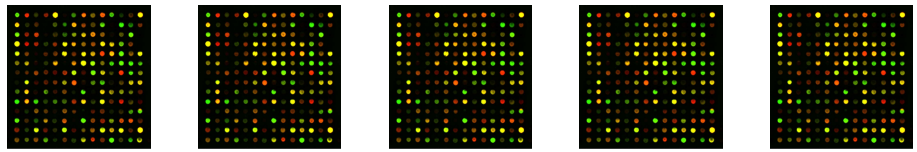


Input : perturbation to the system (e.g. gene overexpression)
Output : measure response to perturbation

Expression data is used to infer the network



A weight is learned for each edge



$\Downarrow$

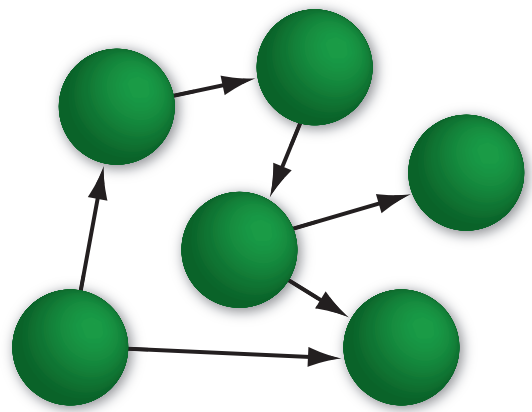
		Target gene			
		gene 1	**gene 2**	$\cdots$	**gene p**
Regulating gene	**gene 1**	-	0.05	$\cdots$	0.56
	gene 2	0.19	-	$\cdots$	0.03
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	gene p	0.11	0.42	$\cdots$	-

Main idea

Tree-based methods

Network inference : GENIE3

Results

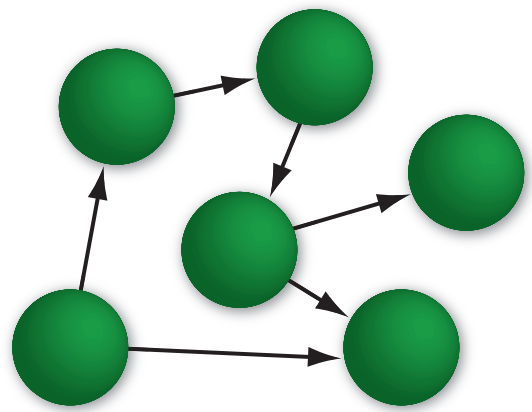


Main idea

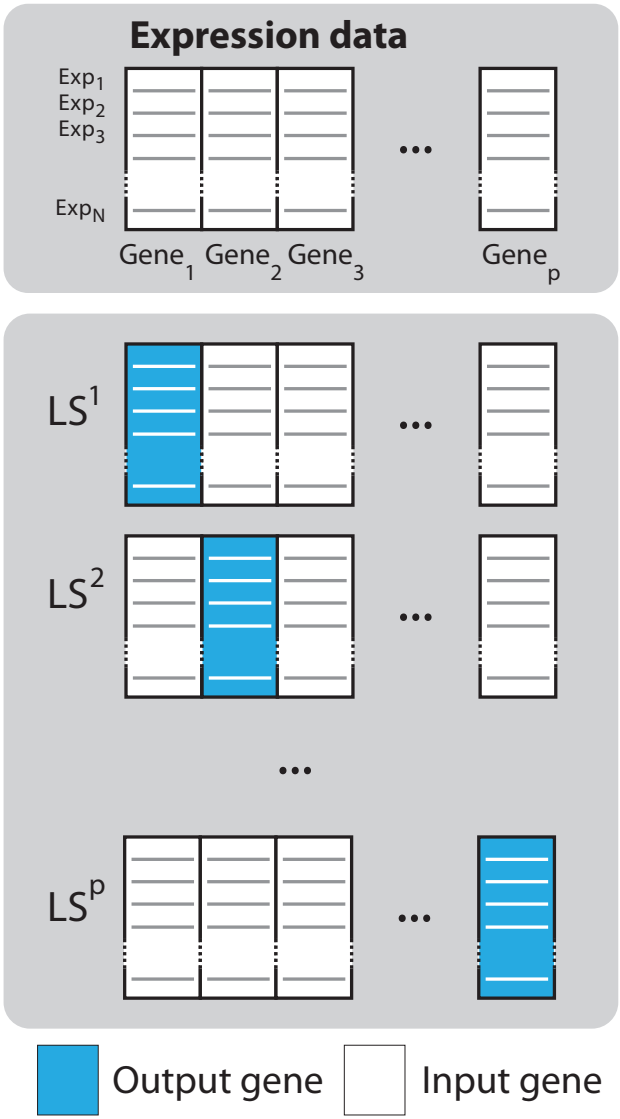
Tree-based methods

Network inference : GENIE3

Results



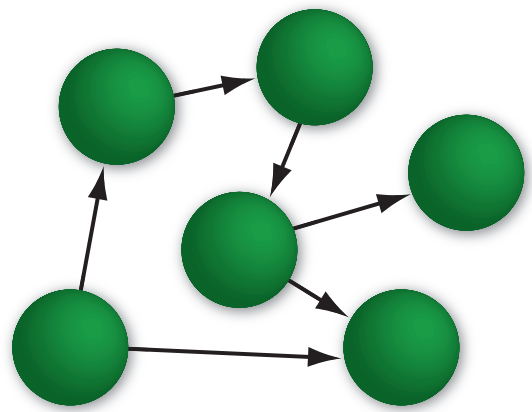
The inference problem decomposes into p sub-problems



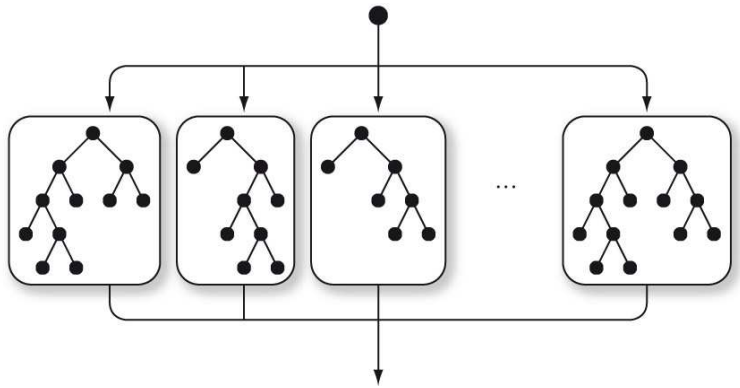
Sub-problem i

$=$

Find the regulators of gene i



Tree-based ensemble methods are good candidates



Bagging
Random Forests
Extra-Trees
...

Non-parametric

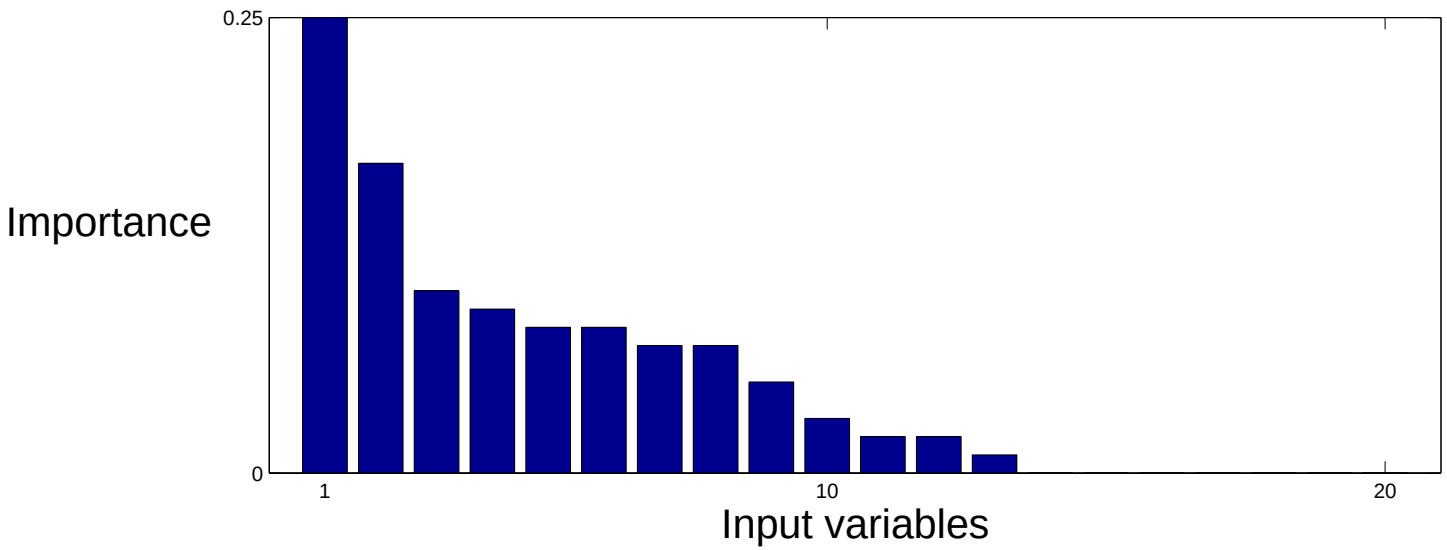
Can deal with interacting features

Work well with high-dimensional datasets

Scalable

The tree-based model is informative

The learned model can be used to find the most relevant inputs.



The variable importance is based on variance reduction

At each tree node $\mathcal{N}$:

$$I(\mathcal{N}) = \#S\mathrm{Var}(S) - \#S_t\mathrm{Var}(S_t) - \#S_f\mathrm{Var}(S_f)$$

S : set of samples reaching node $\mathcal{N}$
S_t (resp. S_f) : subset of S for which the test is true (resp. false)
Var(.) : variance of output variable in a subset

For a single tree :

w_i^t = sum of I at each node where variable i appears

For an ensemble of trees :

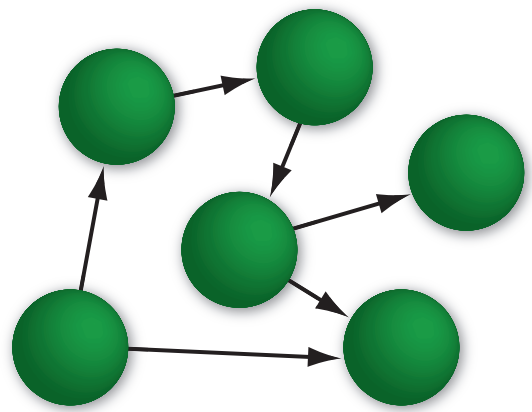
$$w_i = \frac{1}{T}\sum_{t=1}^{T} w_i^t$$

Main idea

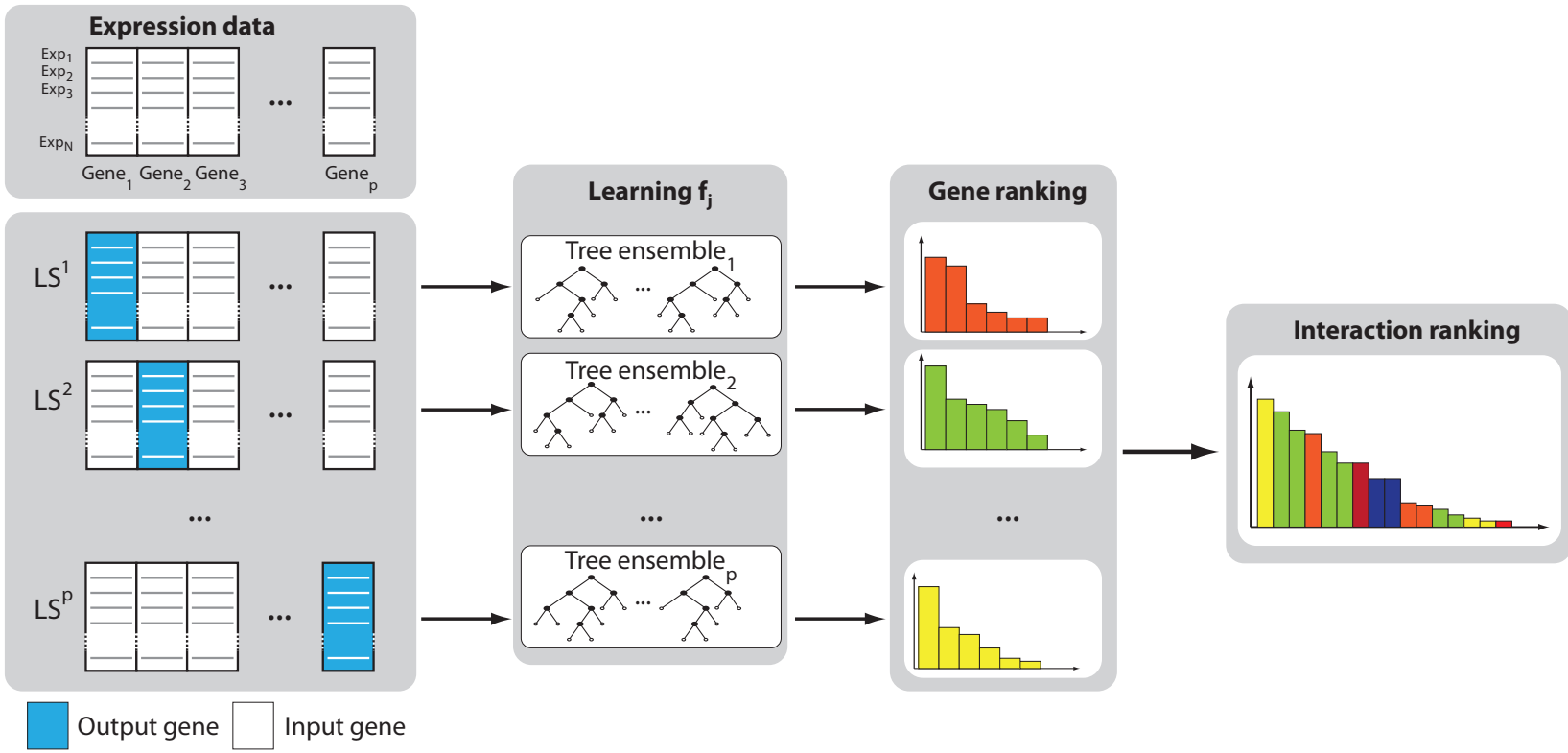
Tree-based methods

Network inference : GENIE3

Results



GENIE3 uses ensembles of trees to infer a network



A normalization is required

For an unpruned tree :

$$\sum_{i \neq j} w_{i \to j} \approx N \mathrm{Var}_j(S)$$

$w_{i \to j}$: importance of gene i for the prediction of gene j
N : number of experiments

$\mathrm{Var}_j(S)$: variance of gene j in the learning sample from which the tree is built

↓

Positive bias for edges towards highly variable genes

↓

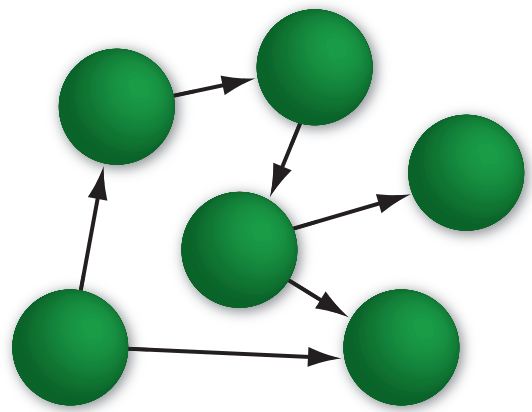
Normalization of gene expressions
so that they have unit variance

Main idea

Tree-based methods

Network inference : GENIE3

Results



GENIE3 is best performer in DREAM4 challenge

DREAM4 *In silico Multifactorial network* challenge :
inference of **synthetic** regulatory networks.

5 networks of 100 genes, 100 experiments per network.

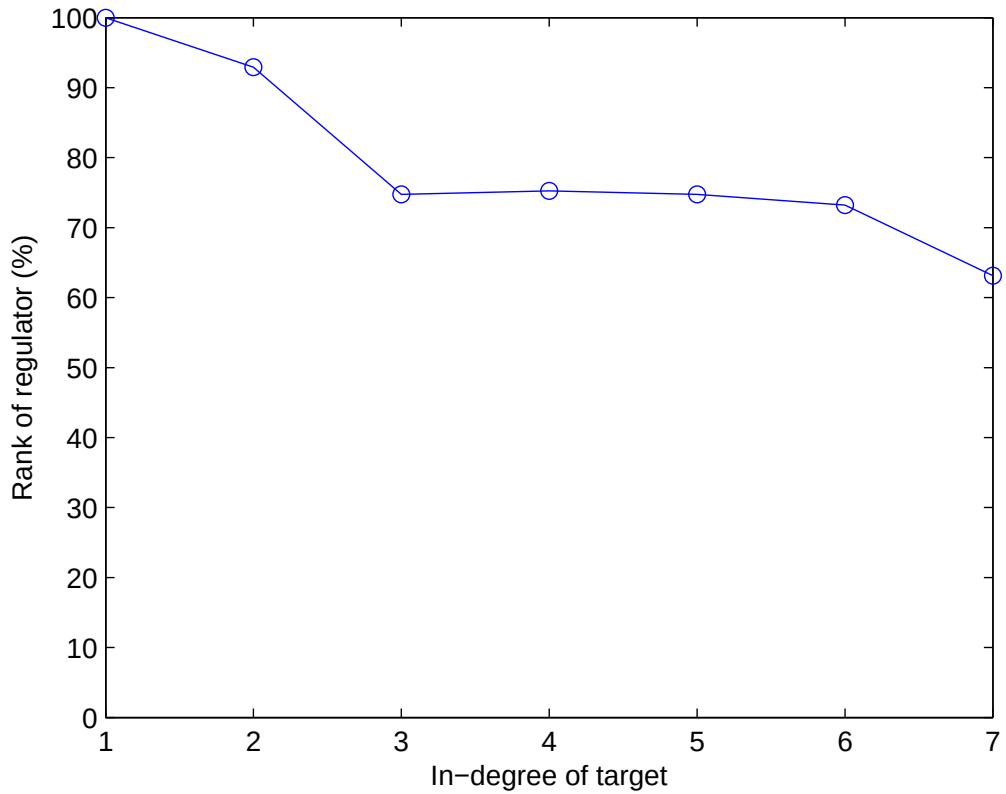
Rank	Team	Mean AUPR	Overall AUPR p-value	Mean AUROC	Overall AUROC p-value
1	GENIE3-Bagging	0.22	5.93e-54	0.76	1.93e-28
2	Team 549	0.14	7.45e-35	0.73	6.29e-23
...	...	...	...	...	...

AUPR : Area Under Precision-Recall curve

AUROC : Area Under ROC curve

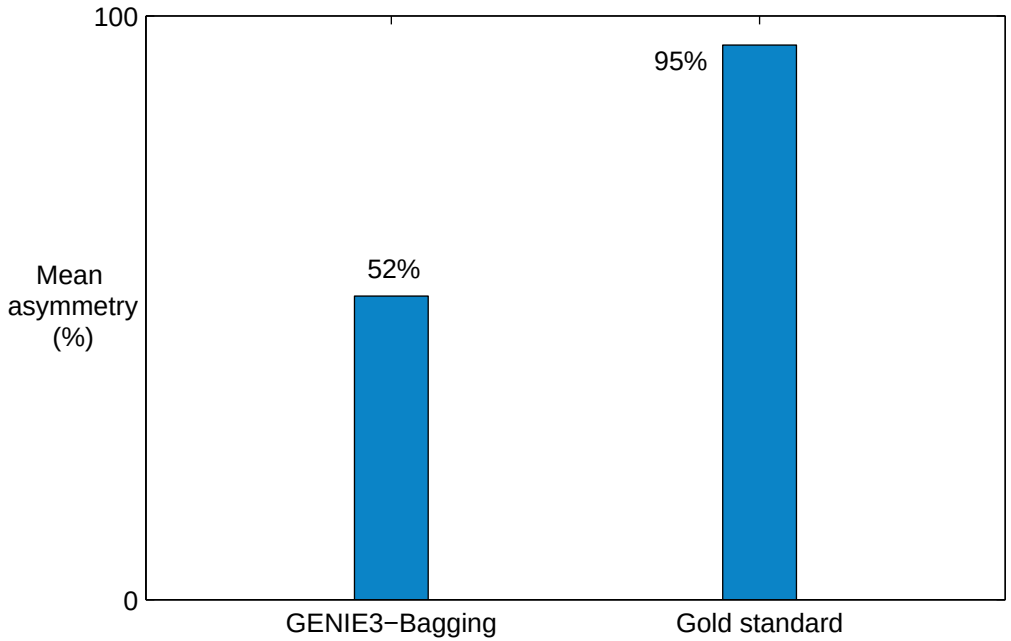
Quality of ranking decreases with in-degree of genes

In-degree = number of regulators



GENIE3 is able to predict a *directed* network

Predicted networks contain a significant number of asymmetric links.



GENIE3 can be used for directing an undirected network

Error on edge directionality :

$i \rightarrow j$ present in gold standard

$j \rightarrow i$ not present in gold standard

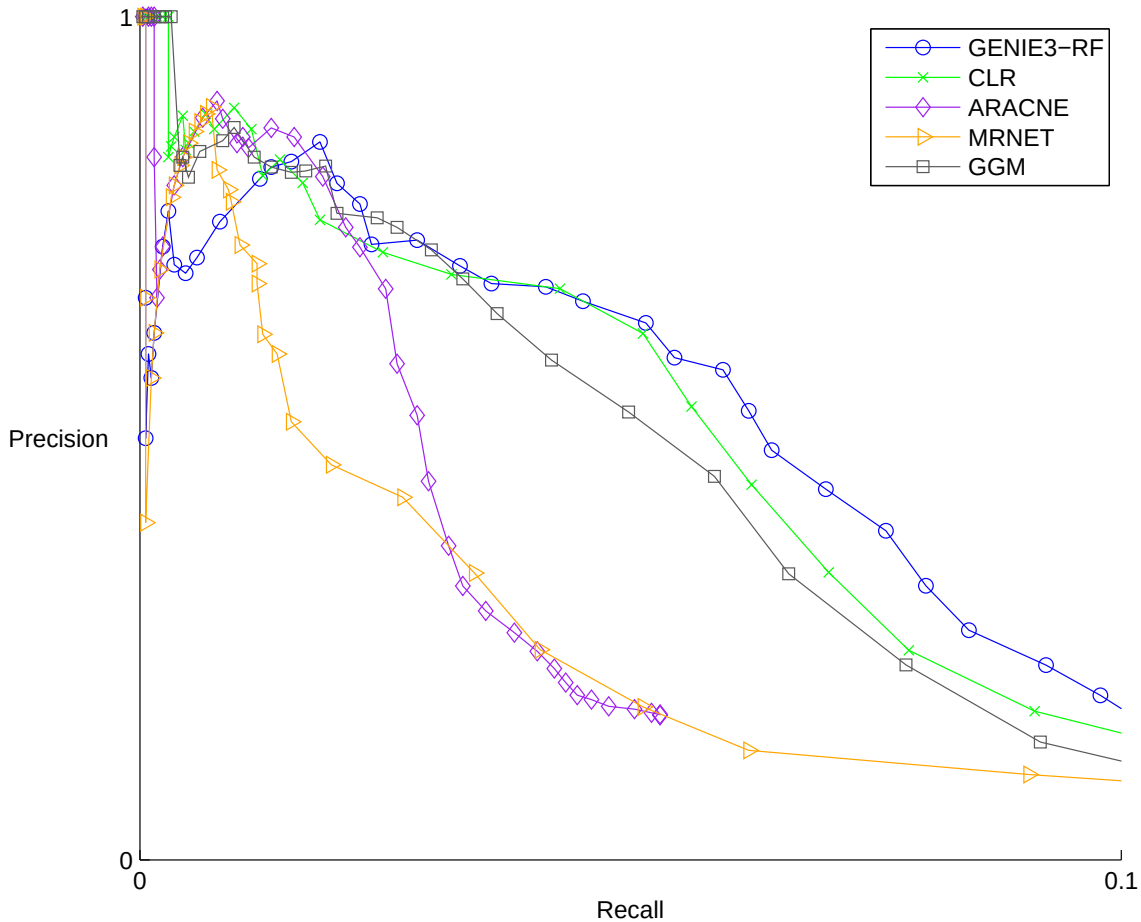
$$w_{i \rightarrow j} < w_{j \rightarrow i}$$

At 5% recall, mean error rate on edge directionality is 20%.

Results on *E. coli* are competitive to existing approaches

1471 genes, 907 experiments. Validation with RegulonDB.

Input genes = 172 known TFs

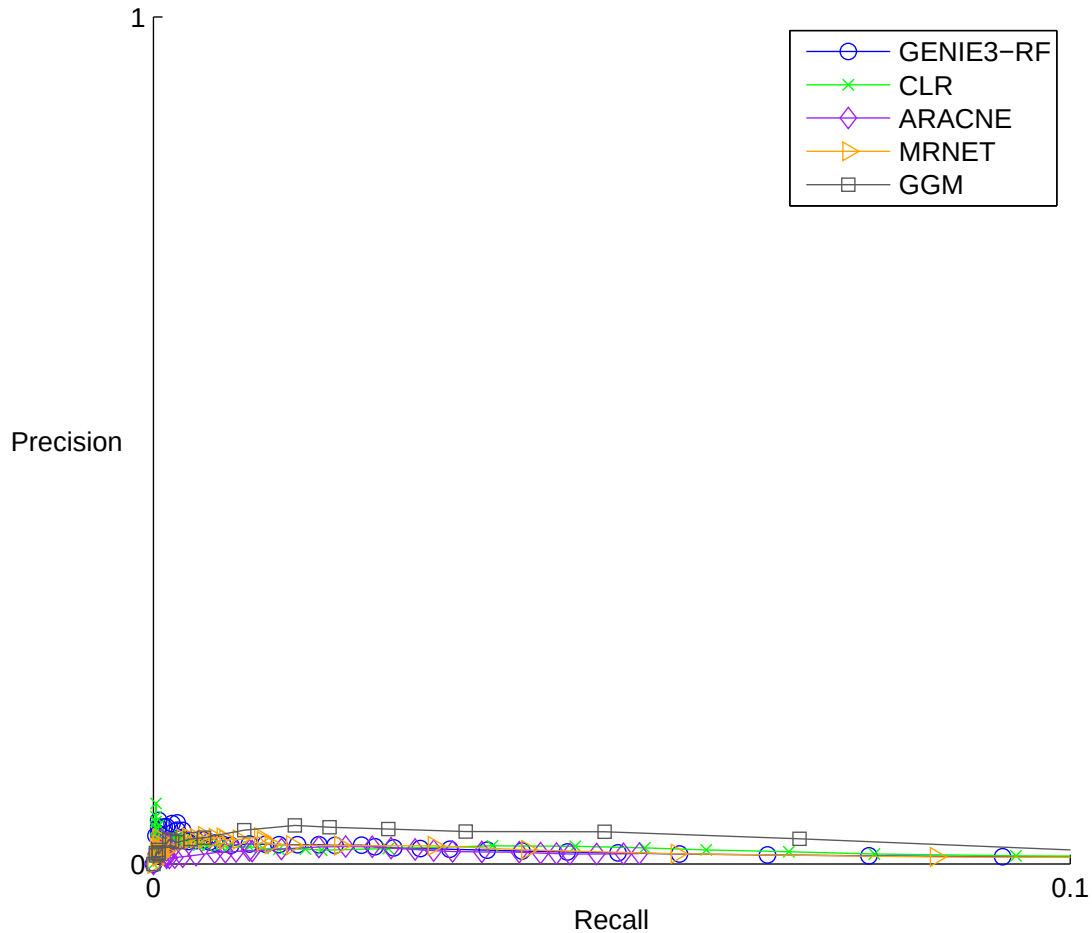


CLR : Faith et al. (2007)
MRnet : Meyer et al. (2007)
ARACNE : Margolin et al. (2006)
GGM : Schafer et al. (2005)

Results on *E. coli* are competitive to existing approaches

1471 genes, 907 experiments. Validation with RegulonDB.

Input genes = all genes



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

Conclusions

Good results for a non parametric approach

Scalable

Can be easily parallelized

Adaptable to other types of genomic data and interactions

Future works

Improvement on the way variable importance scores are normalized

Threshold on the ranking of interactions

Comparison with Bayesian networks

Software :
http ://www.montefiore.ulg.ac.be/∼huynh-thu/software.html

Related paper :
Inferring regulatory networks from expression data using tree-based methods.
Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts
PLoS ONE **5**(9) :e12776