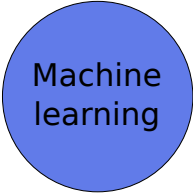
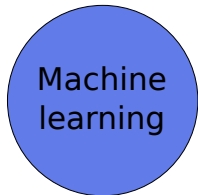

Machine learning-based feature ranking: Statistical interpretation and gene network inference

Vân Anh Huynh-Thu

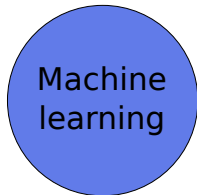
PhD Defense
January 9th, 2011



Machine
learning



Mass spectrometry data analysis
Microarray image analysis
Protein-protein interaction prediction
Coding region identification
Biomarker discovery
Gene function prediction
Gene annotation
Microarray data pre-processing
Splice site detection
SNP analysis
Promoter binding sites identification
Protein function prediction
Gene network inference
Protein annotation
Protein structure prediction
Microarray data analysis



Mass spectrometry data analysis

Microarray image analysis

Protein-protein interaction prediction

Coding region identification

Biomarker discovery

Gene function prediction

Gene annotation

Microarray data pre-processing

Splice site detection

SNP analysis

Promoter binding sites identification

Protein function prediction

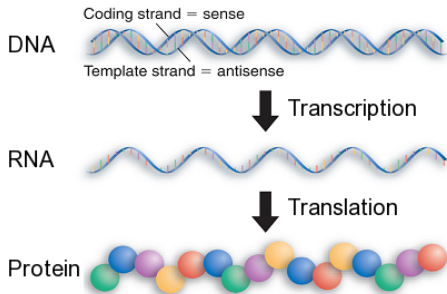
Gene network inference

Protein annotation

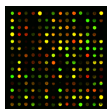
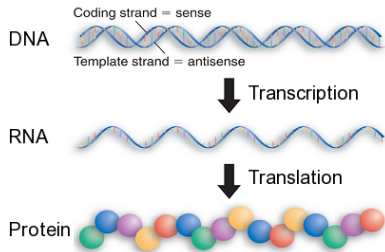
Protein structure prediction

Microarray data analysis

Central dogma of molecular biology



Microarrays measure gene expression levels in a condition

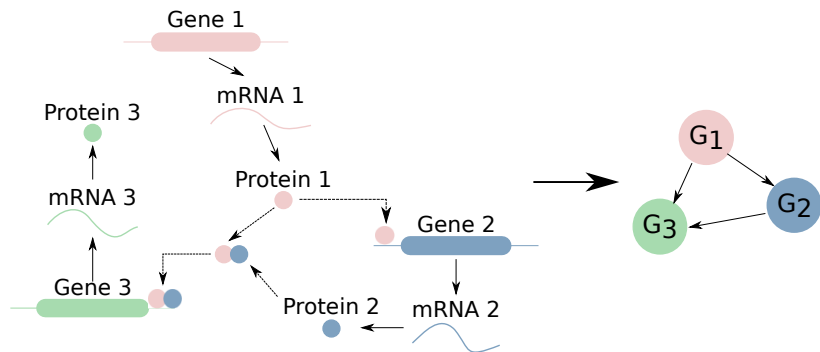


~ 10000 genes

X_1	X_2	...	X_m
-0.61	0.41	...	0.51
-2.3	0.1	...	-0.21
0.33	-0.45	...	0.3
0.23	0.87	...	0.09
...
-0.69	-0.61	...	0.02

} ~ 100 cond.

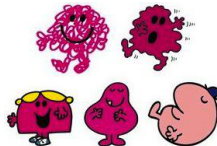
Genes regulate each other



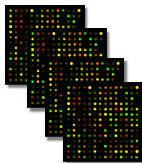
Expression data is used to identify biomarkers



Healthy



Sick

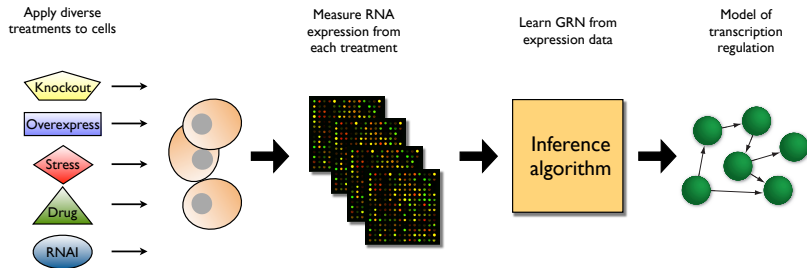


X_1	X_2	...	X_m	Class
-0.61	0.41	...	0.51	Healthy
-2.30	0.10	...	-0.21	Healthy
...
0.33	-0.45	...	0.09	Sick
-0.69	-0.61	...	0.30	Sick



BIOMARKERS

Expression data is used to infer networks



(Gardner & Faith, *Phys Life Rev.*, 2005)

Supervised learning consists in extracting knowledge from input-output pairs

X_1	X_2	\dots	X_m	Y
-0.61	0.41	\dots	0.51	0.56
-2.3	0.1	\dots	-0.21	0.43
0.33	-0.45	\dots	0.3	-0.16
0.23	0.87	\dots	0.09	0.71
\dots	\dots	\dots	\dots	\dots
-0.69	-0.61	\dots	0.02	-0.75



Model

Feature selection/ranking

$$\hat{Y} = f(X_1, X_2, \dots, X_m)$$

feat.	score
X_5	0.248
X_8	0.122
X_2	0.082
\dots	\dots
X_3	0.011

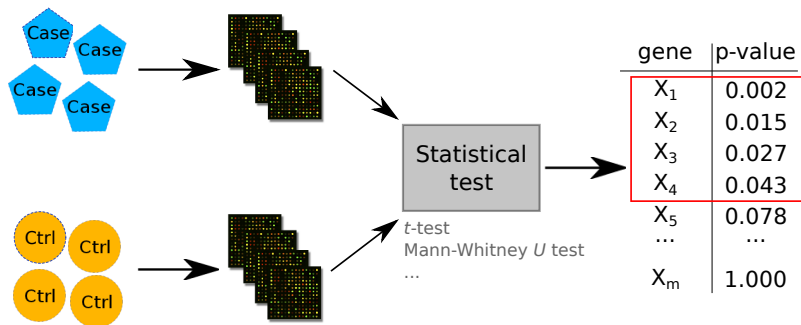
Biomarker discovery and network inference can be treated as feature selection problems

Biomarker discovery: identification of the variables that provide information about a phenotype

Gene network inference: identification of the regulators of each target gene

PART I: BIOMARKER DISCOVERY

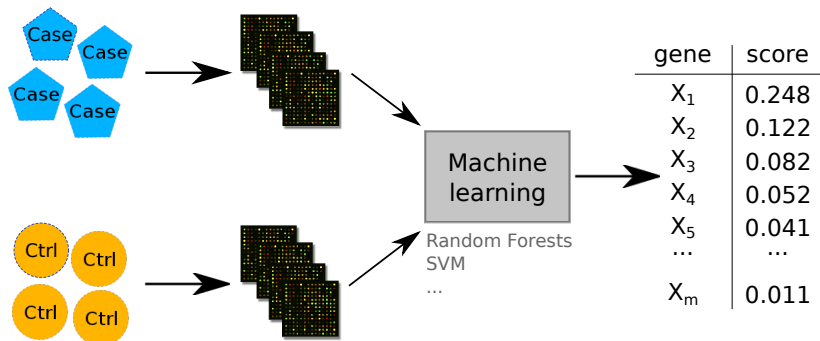
Statistical tests are widely used for biomarker discovery



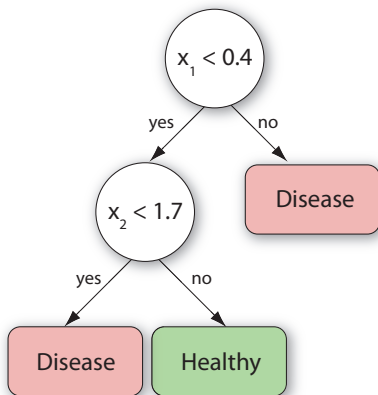
Drawback:

Potentially miss biomarkers that are only relevant in **interaction** with others.

Machine learning methods can deal with interacting features



Decision tree is a supervised learning method



Each interior node tests an input variable.

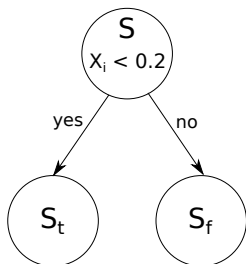
Each leaf node contains a prediction for the output.

The tree-based relevance score is based on entropy reduction

At each tree node:

$$I = \#S.H_Y(S) - \#S_t.H_Y(S_t) - \#S_f.H_Y(S_f)$$

$H_Y(\cdot)$: entropy of the class frequencies in a subset



Relevance score of X_i :

s_i = sum of I at each node where variable X_i appears

Machine learning-based relevance scores are not statistically interpretable

variable	score
X_1	0.248
X_2	0.122
X_3	0.082
X_4	0.052
X_5	0.041
...	...
X_m	0.011

?

Difficult to select a relevance threshold

Prevents the wide adoption
of machine learning methods
by biologists

We seek procedures for extracting features from a ranking

variable	score
X_1	S_1
X_2	S_2
X_3	S_3
...	...
X_{k-1}	S_{k-1}
X_k	S_k
X_{k+1}	S_{k+1}
...	...
X_m	S_m

Select the k top-ranked variables such that:

They contain the highest possible number of relevant variables.

The rate of false positives is as small as possible.

$$s_1 \geq s_2 \geq \dots \geq s_m$$

The original score is replaced by
a statistically interpretable measure

variable	score	FDR
X_1	s_1	0.001
X_2	s_2	0.014
X_3	s_3	0.035
X_4	s_4	0.063
X_5	s_5	0.068
...
X_m	s_m	0.997

$$s_1 \geq s_2 \geq \dots \geq s_m$$

New measure: FDR, FWER, p-value...

The new measure can be interpreted
in a statistical way.

It allows the user to determine a
relevance threshold in a more informed
way.

Feature selection: contributions

Procedure that estimates the **probability to have at least one irrelevant feature** in a subset of top-ranked variables.

(Huynh-Thu *et al.*, *JMLR: Workshop and Conference proceedings*, 2008)

Large-scale evaluation of several procedures for extracting relevant features from a ranking.

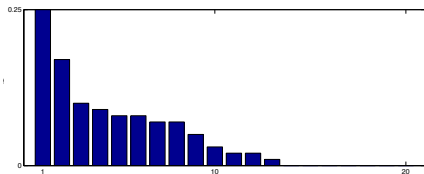
(Huynh-Thu *et al.*, 2012, Submitted)

Conditional error rate

Large-scale evaluation: Methods

Artificial datasets

Microarray datasets

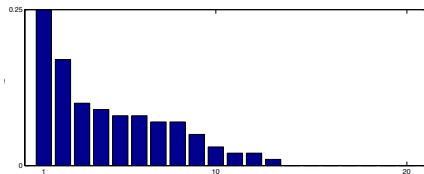


Conditional error rate

Large-scale evaluation: Methods

Artificial datasets

Microarray datasets



Conditional Error Rate

Feat.	Score
X_1	s_1
X_2	s_2
X_3	s_3
...	...
X_{i-1}	s_{i-1}
X_i	s_i
...	...
X_m	s_m

$$s_1 \geq s_2 \geq \dots \geq s_m$$

$$\text{CER}(s_i) = P(\max_{j=i, \dots, m} S_j \geq s_i | H_R^{1 \rightarrow i-1}, H_I^{i \rightarrow m})$$

$H_R^{1 \rightarrow i-1}$: $X_1 \rightarrow X_{i-1}$ relevant

$H_I^{i \rightarrow m}$: $X_i \rightarrow X_m$ irrelevant.

If s_j are univariate statistics:

→ CER = Westfall & Young's maxT adjusted p -value

The CER is estimated by random permutations

var.	score	
X_1	s_1	CER_1
X_2	s_2	CER_2
\dots	\dots	\dots
X_{i-1}	s_{i-1}	CER_{i-1}
X_i	s_i	CER_i
\dots	\dots	\dots
X_m	s_m	CER_m

$$s_1 \geq s_2 \geq \dots \geq s_m$$

For $i = 1, \dots, m$:

- 1 For $p = 1, \dots, P$ (typically $P = 1000$):
 - Keep values of the output and of $X_1 \rightarrow X_{i-1}$ fixed.
 - Randomly (jointly) permute values of $X_i \rightarrow X_m$.
 - Compute variable relevance scores $\{s_1^p, s_2^p, \dots, s_m^p\}$ from permuted data.

2

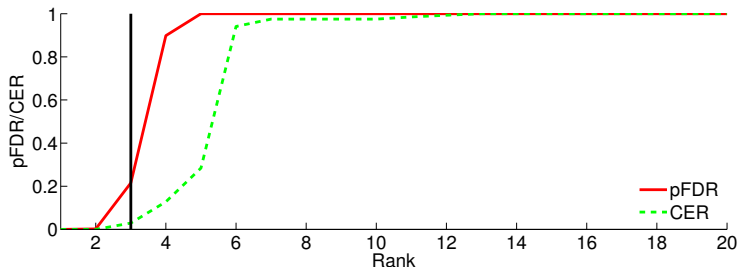
$$\text{CER}_i = \frac{1}{P} \cdot \#\{p : \max_{j=i, \dots, m} s_j^p \geq s_i\}$$

The CER provides a robust estimation of the FWER

Non-linear problem (20 variables, 200 instances)

3 relevant variables

Machine learning algorithm = Extra-Trees



$\text{CER} < 0.05 \rightarrow 3$ variables

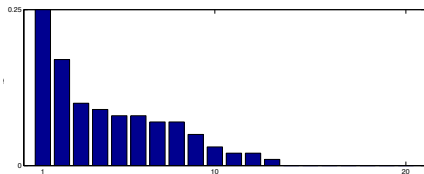
$\text{pFDR} < 0.05 \rightarrow 2$ variables

Conditional error rate

Large-scale evaluation: Methods

Artificial datasets

Microarray datasets



Evaluated methods

Estimation of the generalization error of a model ($\text{err-}\mathcal{A}$, err-TRT)

Multiple testing with random permutations (pFDR , eFDR , CER)

Estimation of the null rank distribution (mr-test)

Introduction of random probes (1Probe , mProbes)

Evaluated methods

Estimation of the generalization error of a model ($\text{err-}\mathcal{A}$, err-TRT)

Multiple testing with random permutations (pFDR , eFDR , CER)

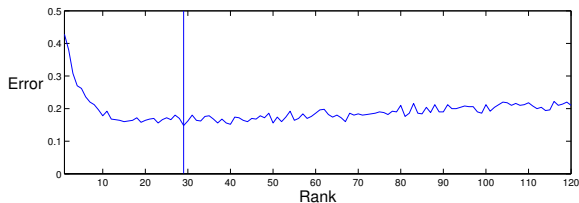
Estimation of the null rank distribution (mr-test)

Introduction of random probes (1Probe , mProbes)

New measure: generalization error e_i of a predictive model \mathcal{A} that uses only the i top-ranked variables (estimated using cross-validation).

Select k top-ranked variables such that:

$$k = \arg \min_{i=1, \dots, m} e_i$$



(see e.g. Geurts *et al.*, *Bioinformatics*, 2005)

1Probe

New measure: probability p_i for a random probe to be ranked above X_i .

variable	score	p-value
X_1	s_1	p_1
X_2	s_2	p_2
X_3	s_3	p_3
...
X_{i-1}	s_{i-1}	p_{i-1}
X_i	s_i	p_i
...
X_m	s_m	p_m

$$s_1 \geq s_2 \geq \dots \geq s_m$$

- 1 For $p = 1, \dots, P$ (typically $P = 1000$):
 - Create a random feature X_{rand} (e.g. $\sim \mathcal{N}(0, 1)$).
 - Add X_{rand} to the original dataset.
 - Compute variable relevance scores $\{s_1^p, \dots, s_m^p, s_{rand}^p\}$ from new dataset.
- 2 Proportion of runs where X_{rand} is ranked above X_i :

$$p_i = \frac{1}{P} \cdot \#\{p : s_{rand}^p \geq s_i^p\}$$

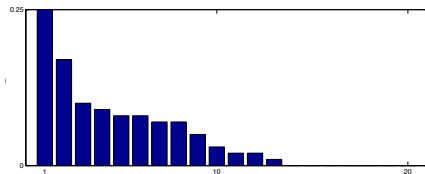
(inspired from Stoppiglia *et al.*, *JMLR*, 2003)

Conditional error rate

Large-scale evaluation: Methods

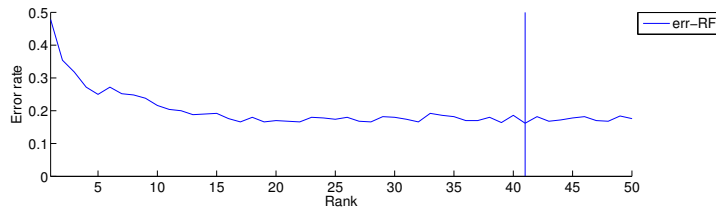
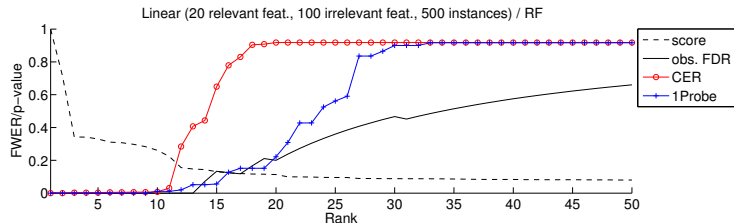
Artificial datasets

Microarray datasets



Most methods improve the interpretability of the original relevance score

Machine learning algorithm = Random Forests



Precision and recall are used as performance metrics

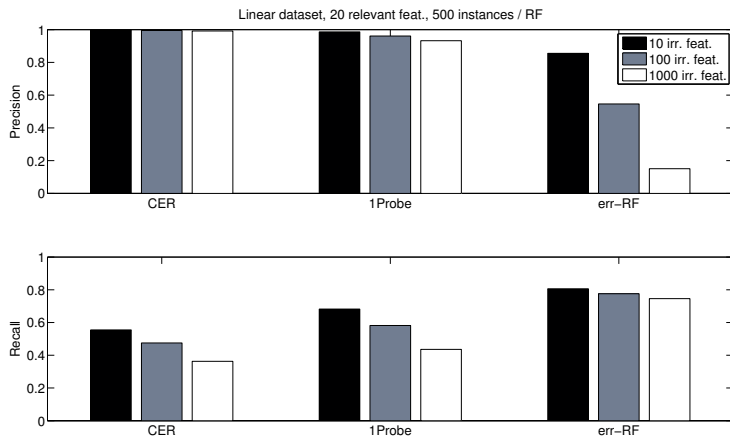
variable	score	CER	
X_1	S_1	0.00	relevant
X_2	S_2	0.01	relevant
X_3	S_3	0.02	irrelevant
...
X_S	S_S	0.04	relevant
X_{S+1}	S_{S+1}	0.07	relevant
...
X_m	S_m	1.00	irrelevant

TP: number of selected variables that are relevant

$$\text{precision} = \frac{TP}{S}$$

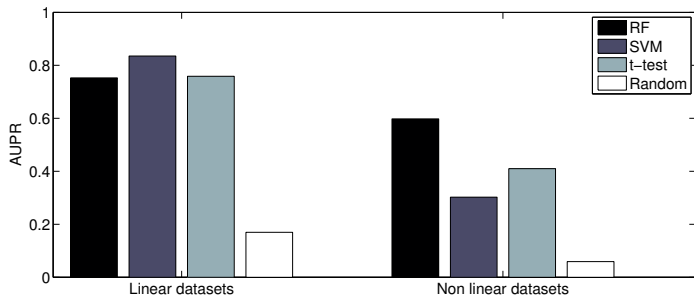
$$\text{recall} = \frac{TP}{\#\text{relevant}}$$

The methods differ in terms of false positives/negatives



(average over 50 datasets in each case)

Comparison of ranking methods



(average over 50 datasets in each case)

Conditional error rate

Large-scale evaluation: Methods

Artificial datasets

Microarray datasets

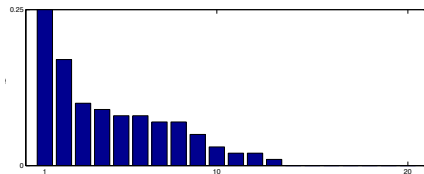
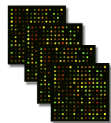


Illustration on a microarray dataset

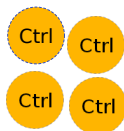
patients with
prostate cancer



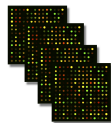
34 samples



healthy patients



19 samples



~ 4000 genes

(Dhanasekaran *et al.*, *Nature*, 2001)

The number of selected genes depends on the ranking method and the values of its parameters

	RF (1000 trees)	RF (10000 trees)	SVM	<i>t</i> -test
CER	58	136	0	391
1Probe	54	444	2	1608
err- \mathcal{A}	5	4	8	–

Part I: Summary

We proposed a procedure (CER) for extracting relevant variables from a ranking derived from a multivariate approach.

The CER procedure takes into account the dependencies between the relevance scores.

We performed a large-scale evaluation of the CER procedure and of other methods that replace the relevance score with a statistically interpretable measure.

The choice of a method depends on the FP/FN tradeoff one wants to achieve.

Selecting the k top-ranked features minimizing some cross-validated error is counter-productive.

Future research directions

Use the number of selected variables as a criterion to tune the parameters of a ranking algorithm (instead of prediction performance).

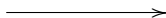
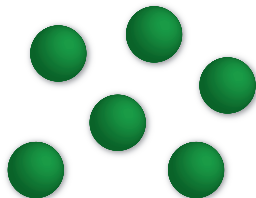
Adapt the procedures for the identification of a minimal subset of relevant variables.

Assess the real interest of multivariate approaches for feature ranking.

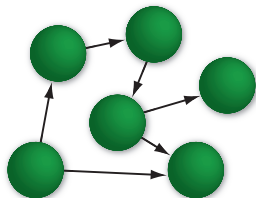
PART II: NETWORK INFERENCE

Inferring regulatory networks is a challenging problem

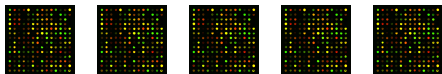
unknown network



inferred network



A weight is learned for each edge



		Target gene			
		gene 1	gene 2	...	gene p
Regulating gene	gene 1	-	0.05	...	0.56
	gene 2	0.19	-	...	0.03

	gene p	0.11	0.42	...	-

Network inference: contributions

Algorithm for the inference of gene regulatory networks from **steady-state expression** data (GENIE3).

(Huynh-Thu *et al.*, *PLoS ONE*, 2010)

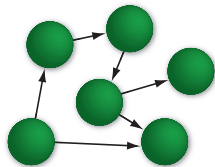
Extensions of GENIE3 to **time series** and to **genetical genomics** data.

Network inference with GENIE3

The DREAM challenges

GENIE3 and time series

GENIE3 and genetical genomics

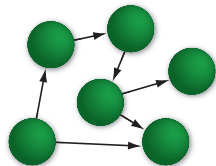


Network inference with GENIE3

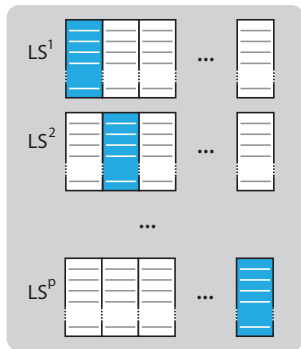
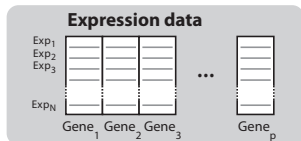
The DREAM challenges

GENIE3 and time series

GENIE3 and genetical genomics



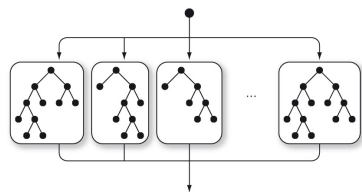
Inference is decomposed into p sub-problems



 Output gene  Input gene

Sub-problem i
=
Find the regulators of gene i

Tree-based ensemble methods are good candidates



Bagging

Random Forests

Extra-Trees

...

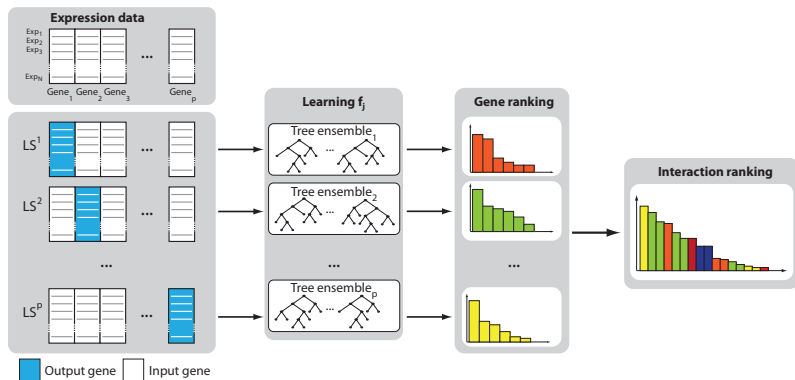
Can deal with interacting features

Non-parametric

Work well with high-dimensional datasets

Scalable

GENIE3: GENE Network Inference with Ensemble of trees

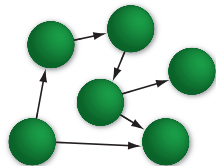


Network inference with GENIE3

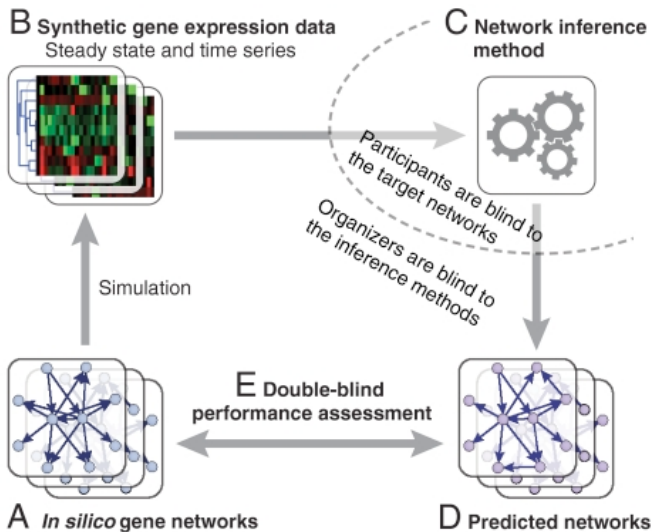
The DREAM challenges

GENIE3 and time series

GENIE3 and genetical genomics

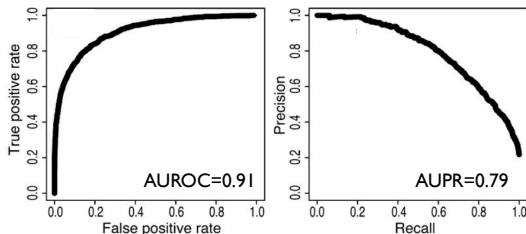


DREAM is an annual reverse engineering competition



(Marbach *et al.*, *PNAS*, 2010)

Evaluation protocol (all challenges)



Output of algorithms: a ranked list of (directed) interactions

Evaluation through ROC and Precision-Recall curves

→ Area under ROC (AUROC) and Precision-Recall (AUPR) curves

→ p-values under random model

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

The DREAM challenges

DREAM4 In Silico Multifactorial:

5 synthetic networks of 100 genes

Data: steady-state expression profiles obtained from slight perturbations of all genes

DREAM5 Network Inference:

1 synthetic and 2 real networks (*E. coli* and *S. cerevisiae*)

Data: microarray compendia

GENIE3 was the overall best performer in both challenges

DREAM4

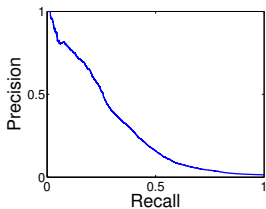
DREAM5

Rank	Team	Overall score
1	GENIE3	37.428
2	Team 549	28.165
3	Team 498	27.053
4	Team 395	26.139
5	Team 425	25.905
...

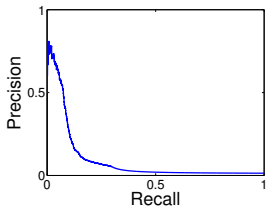
Rank	Team	Overall score
1	GENIE3	40.279
2	Team 543	34.023
3	Team 776	31.099
4	Team 862	28.747
5	Team 548	22.711
...

Methods yield more accurate predictions
for the artificial network than for the real networks

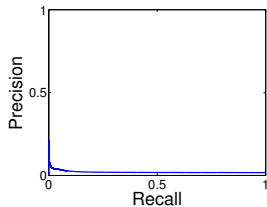
In silico



E. coli



S. Cerevisiae

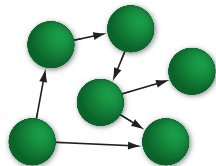


Network inference with GENIE3

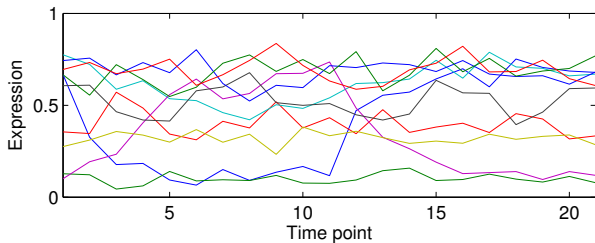
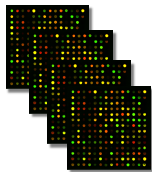
The DREAM challenges

GENIE3 and time series

GENIE3 and genetical genomics



Time series of gene expressions



GENIE3 with time series data

GENIE3-time:

Weight of $g_i \rightarrow g_j$ is the importance of expression of g_i at time t for the prediction of expression of g_j at time $t + h$.

Learning sample:

Inputs				Output
$g_1(t_1)$	$g_2(t_1)$	\dots	$g_p(t_1)$	$g_j(t_1 + h)$
$g_1(t_2)$	$g_2(t_2)$	\dots	$g_p(t_2)$	$g_j(t_2 + h)$
$g_1(t_3)$	$g_2(t_3)$	\dots	$g_p(t_3)$	$g_j(t_3 + h)$
\dots	\dots	\dots	\dots	\dots

GENIE3 with time series + steady-state data

GENIE3-comb:

Learn a **single** model from both datasets by merely concatenating them.

Learning sample:

Inputs				Output
$g_1(exp_1)$	$g_2(exp_1)$	\dots	$g_p(exp_1)$	$g_j(exp_1)$
$g_1(exp_2)$	$g_2(exp_2)$	\dots	$g_p(exp_2)$	$g_j(exp_2)$
$g_1(exp_3)$	$g_2(exp_3)$	\dots	$g_p(exp_3)$	$g_j(exp_3)$
\dots	\dots	\dots	\dots	\dots
$g_1(t_1)$	$g_2(t_1)$	\dots	$g_p(t_1)$	$g_j(t_1 + h)$
$g_1(t_2)$	$g_2(t_2)$	\dots	$g_p(t_2)$	$g_j(t_2 + h)$
$g_1(t_3)$	$g_2(t_3)$	\dots	$g_p(t_3)$	$g_j(t_3 + h)$
\dots	\dots	\dots	\dots	\dots

Weights of edges are averaged over different values of h

If T time points:

$$\begin{array}{cccccccc} h = 1 & h = 2 & h = 3 & h = 4 & h = 5 & \dots & h = T - 1 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow \\ w_{i,j}^1 & w_{i,j}^2 & w_{i,j}^3 & w_{i,j}^4 & w_{i,j}^5 & \dots & w_{i,j}^{T-1} \end{array}$$

Weight for edge $g_i \rightarrow g_j$:

$$w_{i,j} = \frac{1}{T-1} \sum_{h=1}^{T-1} w_{i,j}^h$$

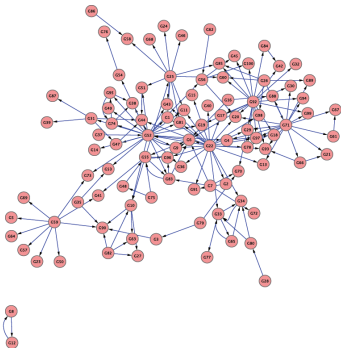
DREAM3 and DREAM4 In Silico Size 100 challenges

Inference of synthetic regulatory networks

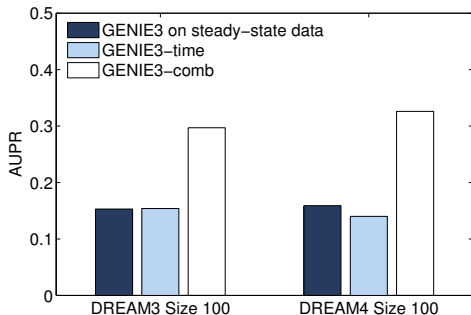
5 networks of 100 genes

Steady-state data:
201 profiles (wild-type, systematic knockout and knockdown of each gene)

Time series:
21 time points in each perturbation experiment



Integrating both types of data improves the predictions



GENIE3-comb would have been ranked:

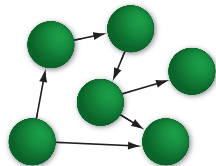
- 2nd on DREAM3
- 3rd on DREAM4

Network inference with GENIE3

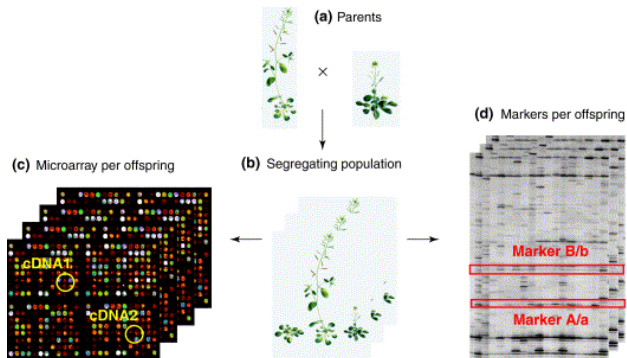
The DREAM challenges

GENIE3 and time series

GENIE3 and genetical genomics



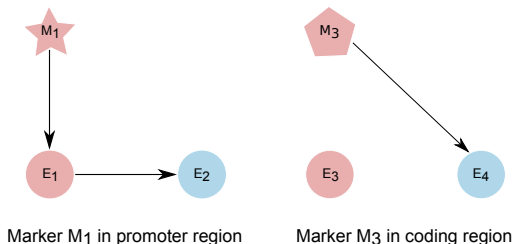
Genetical genomics



TRENDS in Genetics

(Jansen & Nap, *Trends in Genetics*, 2001)

GENIE3 with genetical genomics data



GENIE3-gen-2:

Train two **separate** models from expression and genetic markers

+

Product of markers and expression importance scores

$g_i \rightarrow g_j$ if *both* marker and expression of g_i are predictive of expression of g_j

DREAM5 Systems Genetics challenge

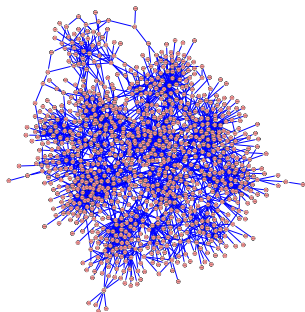
Inference of synthetic regulatory networks

1000 genes in each network

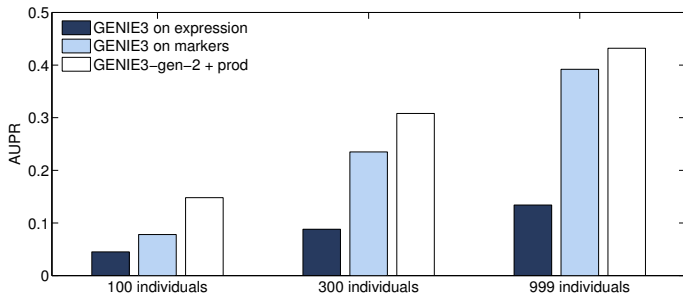
1 marker per gene

Populations of 100, 300, and 999 individuals

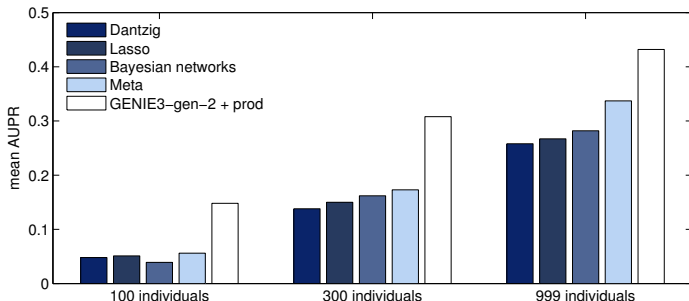
5 networks per population size



Genetic markers bring more information than expression data



GENIE3 outperforms the methods of the best performer



Part II: Summary

GENIE3 yields state-of-the-art performances.

GENIE3 can be extended to different kinds of data and interactions.

Performances of inference methods on real datasets are typically worse than on artificial data.

Future research directions

Application to real datasets

Exploitation of other types of data (e.g. miRNAs)

Extension to the differential networking problem

Supervised inference of regulatory networks

Machine learning-based feature ranking: Statistical interpretation and gene network inference

Vân Anh Huynh-Thu

PhD Defense
January 9th, 2011