# Inferring Gene Regulatory Networks using Ensembles of Feature Selection Techniques

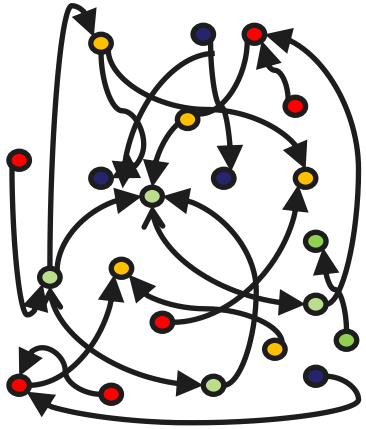
Joeri Ruyssinck, Tom Dhaene & Yvan Saeys

joeri.ruyssinck@intec.ugent.be>
www.ibcn.intec.ugent.be
Internet Based Communication Networks and Services (IBCN)
Department of Information Technology (INTEC)
Ghent University - IBBT

ibbt
connect.innovate.create

UNIVERSITEIT GENT

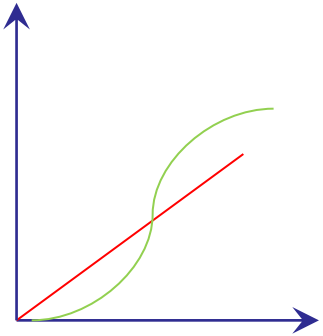
# Presentation flow



Biological problem

| | $g_1$ | ... | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|
| $s_1$ | 0.89 | ... | 0.11 | 0.23 |
| ..... | | | | |
| $s_{p-1}$ | 0.43 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | ... | 0.23 | 0.21 |

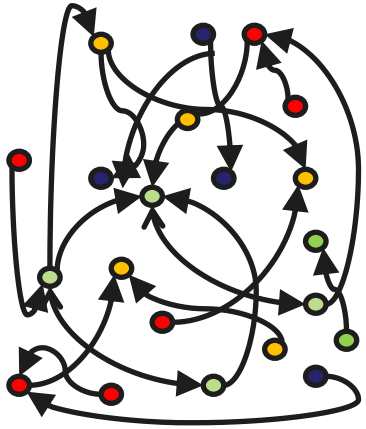
Machine learning problem



Results

Open issues

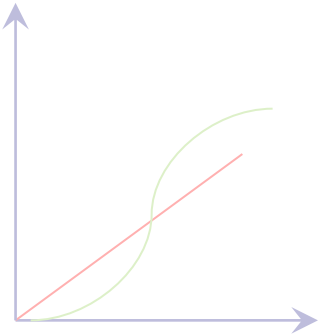
# Biological problem



**Biological problem**

| | $g_1$ | ... | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|
| $s_1$ | 0.89 | ... | 0.11 | 0.23 |
| ... | | | | |
| $s_{p-1}$ | 0.43 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | ... | 0.23 | 0.21 |

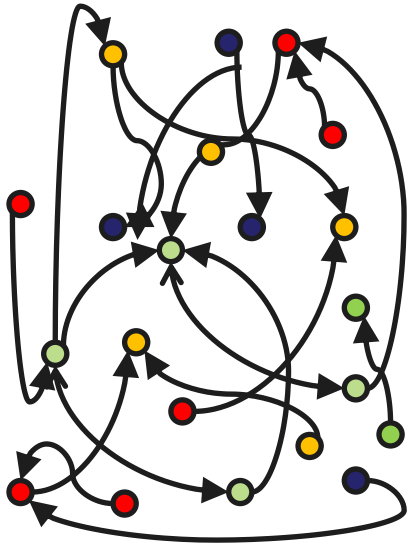
**Machine learning problem**



**Results**

**Open issues**

# What is a gene regulatory network?



= gene

= regulates transcription rate

**Gene regulatory network**

***GRN***

**Network representation of which genes influence the transcription rate of other genes through their expression products**

# Why do we need a gene regulatory network?

Very important **information source**

Explanation of known interactions/processes

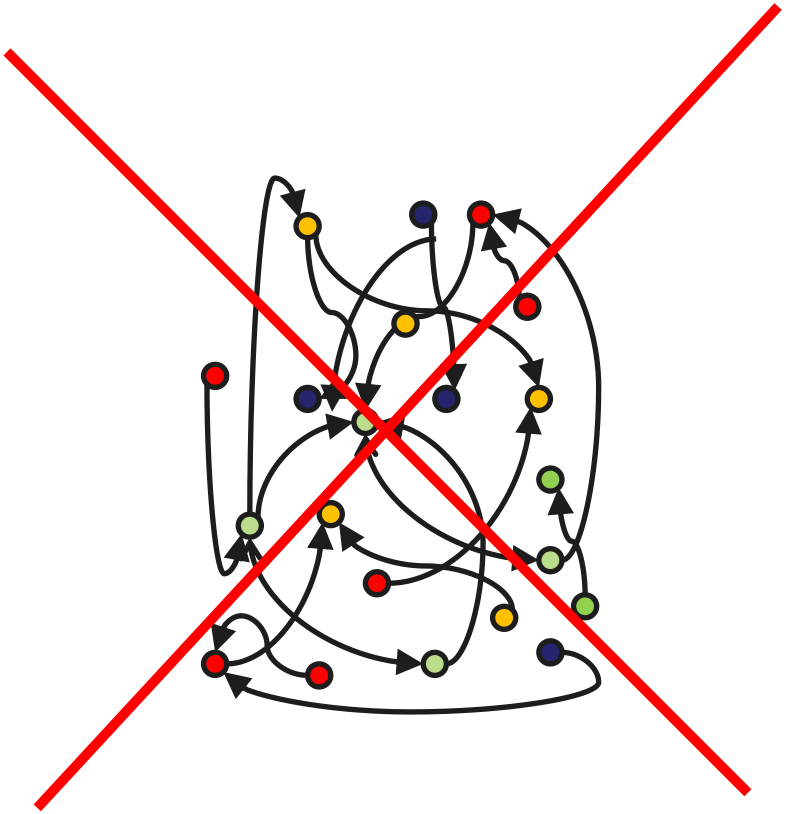
Finding genes potentially related to studied process

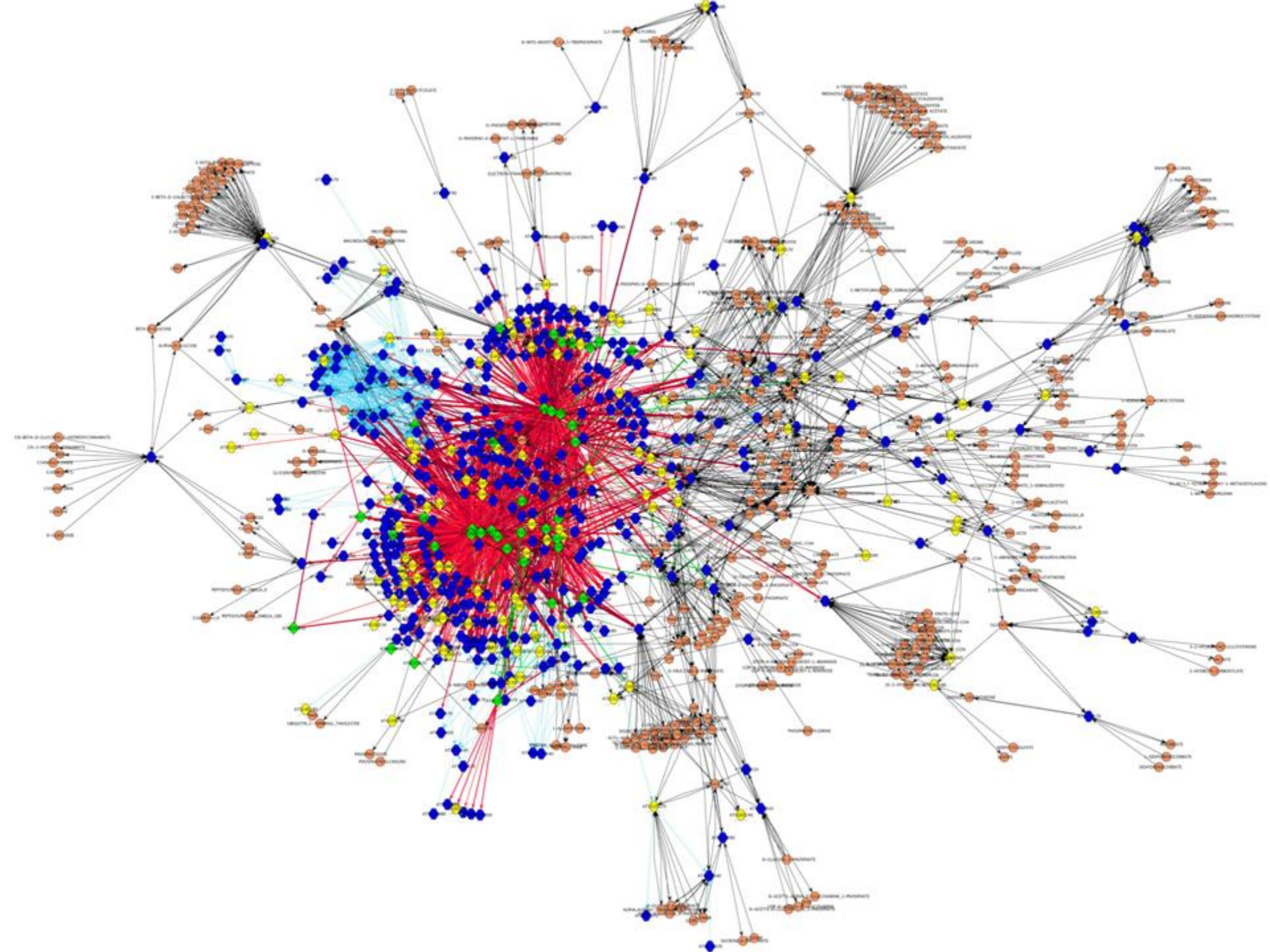
**Leaf growth**

**Pathology / Drug research**

Madhamshettiwar *et al. (*2012) .
Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine*

# Why don’t we have this knowledge?

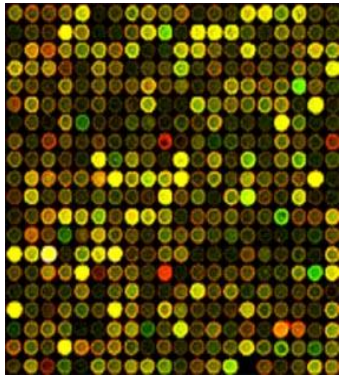




# Biological problem

Too many genes! (20k-25k in humans)

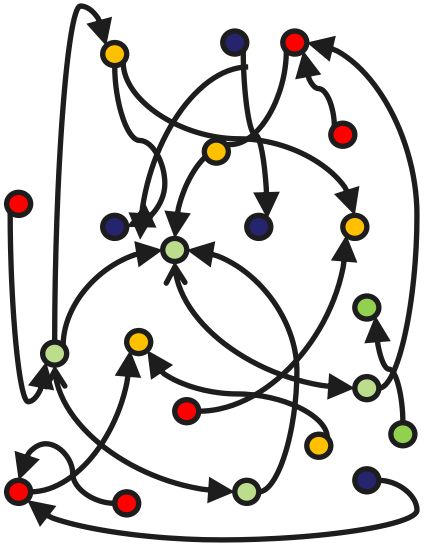
Clear need for high-throughput inference of these gene regulatory networks

Development of computational inference methods which use gene expression data

# Inference problem

| | $g_1$ | $g_2$ | | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|---|
| $s_1$ | 0.89 | 0.45 | ... | 0.11 | 0.23 |
| $s_2$ | 0.11 | 0.83 | ... | 0.32 | 0.44 |
| | ..... | | | | |
| $s_{p-1}$ | 0.43 | 0.22 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | 0.11 | ... | 0.23 | 0.21 |



**Matrix *N X P* with continuous data**

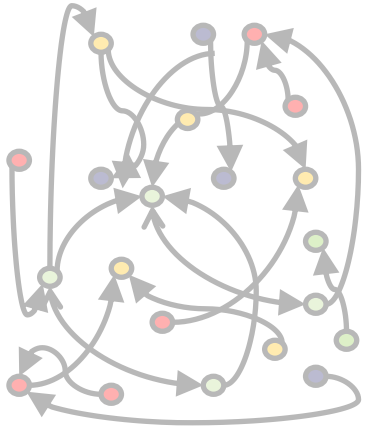
**Gene regulatory network**

---

**Challenges:**

- ***n* >> *p*** **(e.g. 800 samples versus 4000 genes)**
- **Indirect effects**
- **Data is noisy**

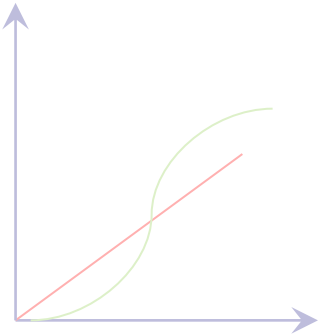
# Machine learning problem



Biological problem

| | $g_1$ | ... | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|
| $s_1$ | 0.89 | ... | 0.11 | 0.23 |
| ..... | | | | |
| $s_{p-1}$ | 0.43 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | ... | 0.23 | 0.21 |

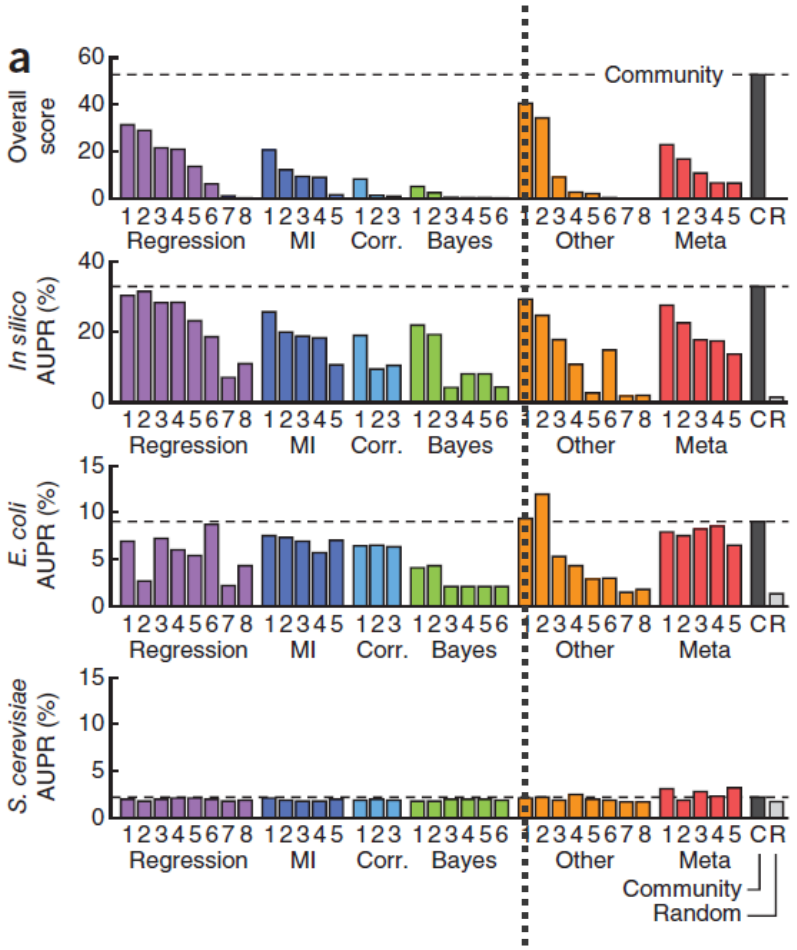
**Machine learning problem**



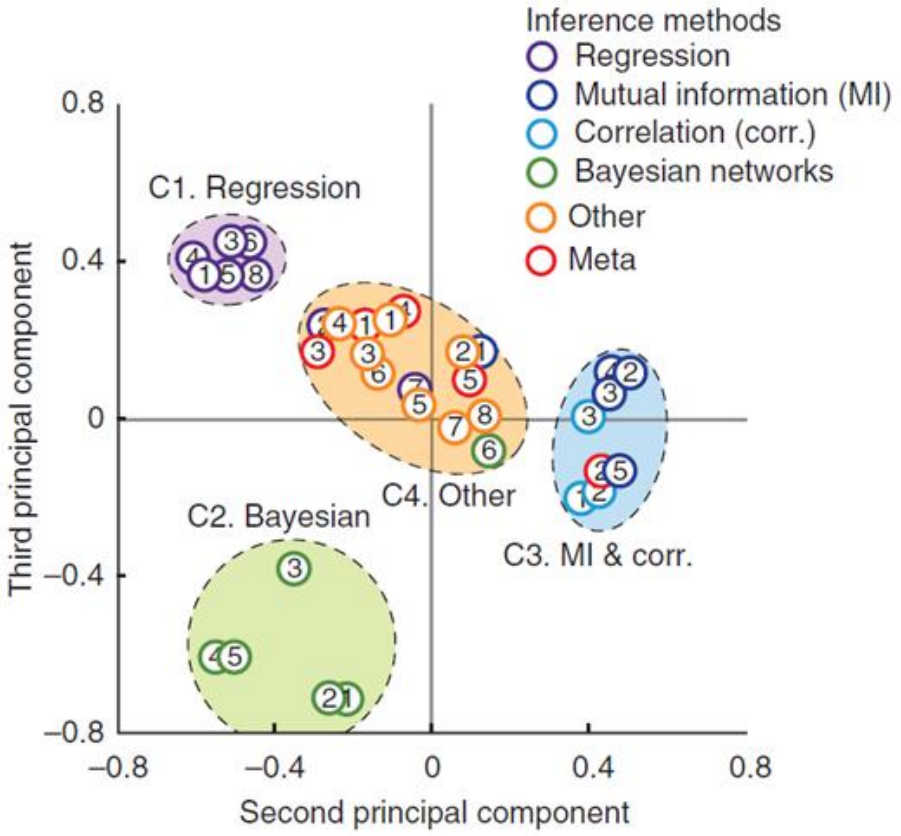
Results

Open issues

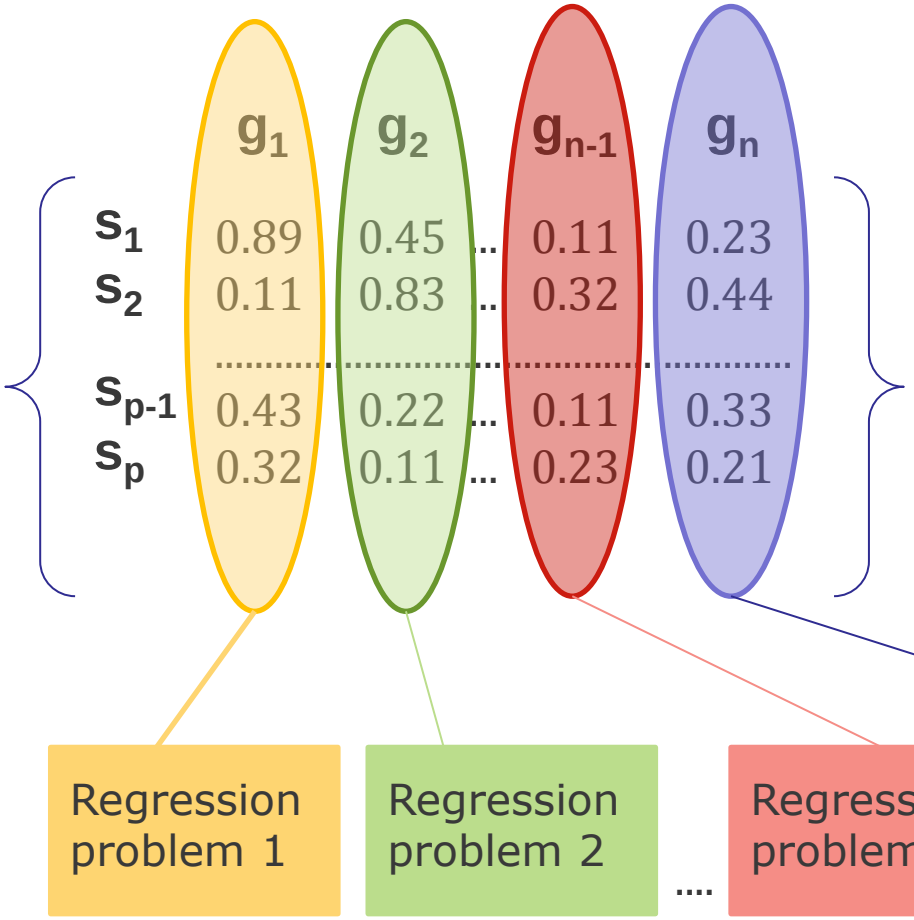
# DREAM5 network inference challenge results




GENIE3



# GENIE3




**Split into *n* regression problems**

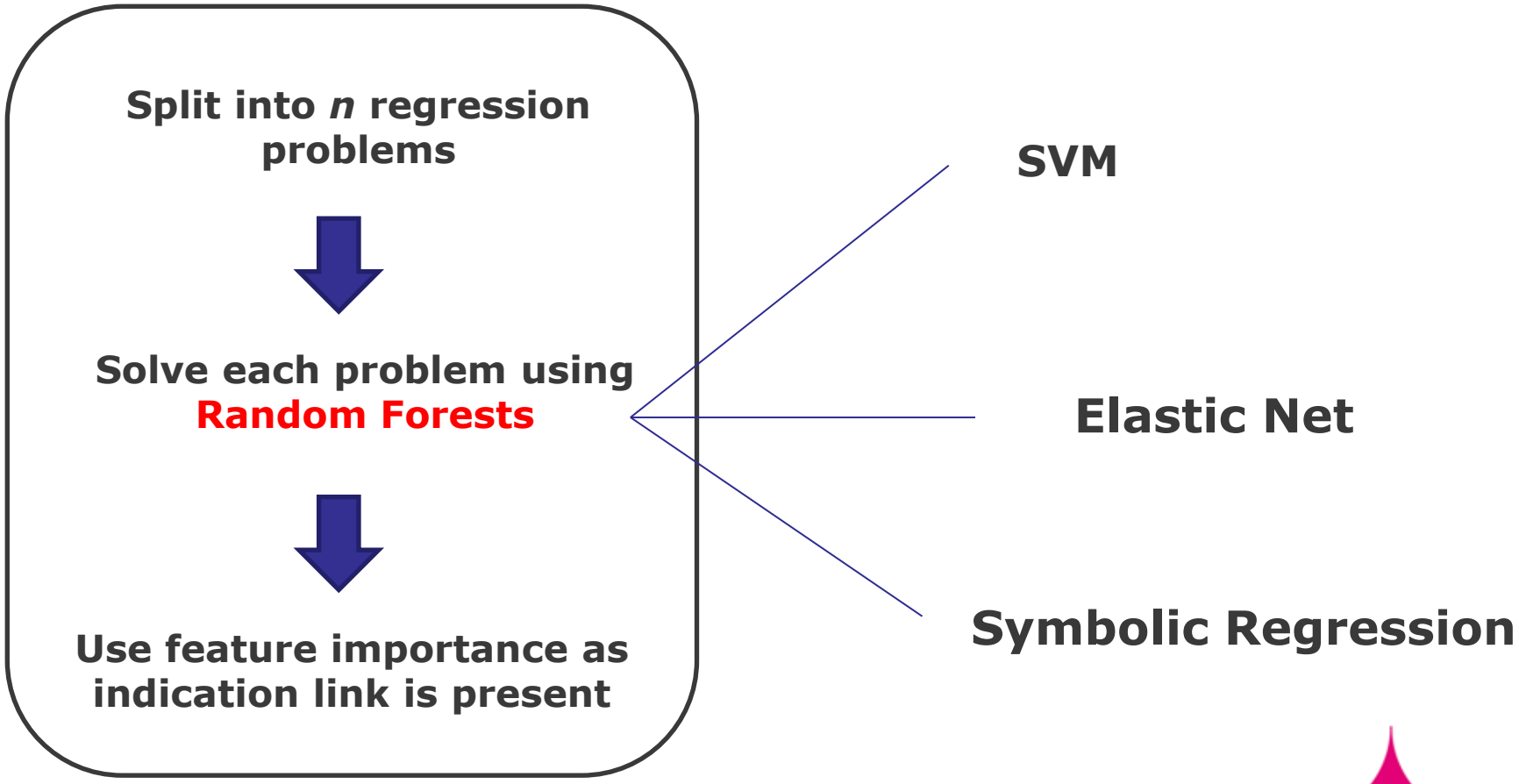
**Model each problem using Random Forests**

**Use feature importance as indication link is present**

**Aggregate scores and rank decreasingly**
**Add links in this order to the network**

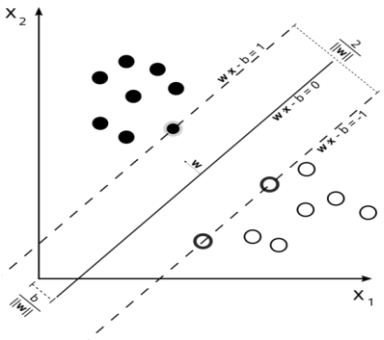
# Question

**Why don't we use other feature selection techniques?**

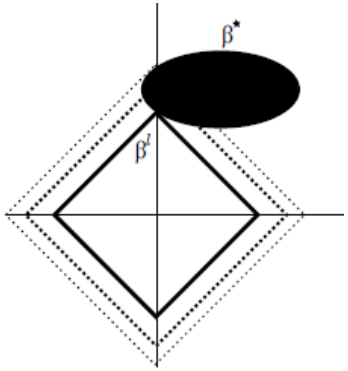


# Feature importance

**SVM**

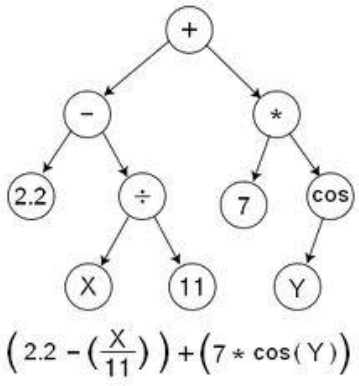


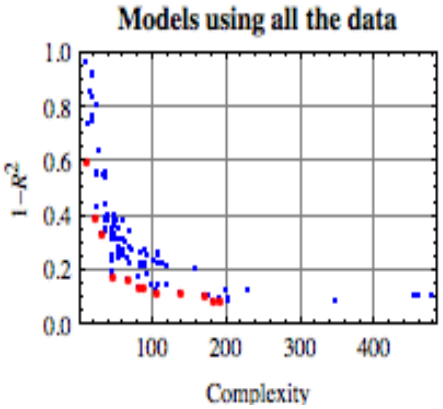

**FI:** absolute value of the feature component in weight vector




**Elastic net**

**FI:** coefficients after a fit with a regularised linear model ($L_1$ / $L_2$ norm)




Find algebraic formula mapping input-output.

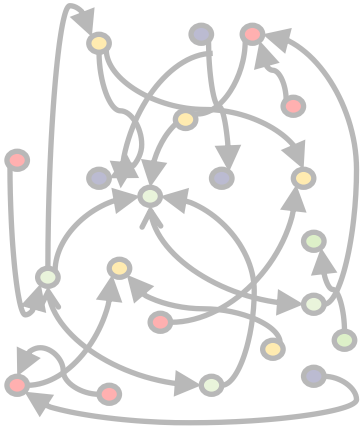
Reduce search space by genetic programming

Optimise two objectives:
goodness of fit and complexity of formula

**FI**: presence of variable in final population

**Symbolic Regression**

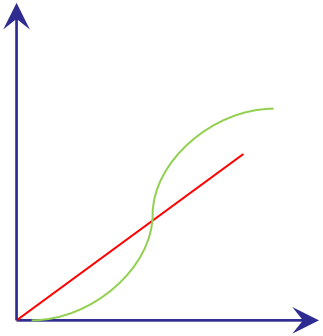
# Results



Biological problem

| | $g_1$ | ... | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|
| $s_1$ | 0.89 | ... | 0.11 | 0.23 |
| ... | | | | |
| $s_{p-1}$ | 0.43 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | ... | 0.23 | 0.21 |

Machine learning problem



Results

Open issues

# DREAM4 dataset

| E.Net. | SVM-lin | SR | GENIE3 | |
|---|---|---|---|---|
| 0.64 | 0.57 | **0.72** | **0.75** | (AUROC) |
| 0.16 | 0.03 | **0.19** | **0.20** | (AUPR) |

Only algorithm achieving similar performance is Symbolic Regression

Similarity between these algorithms: both inherently **ensemble** methods

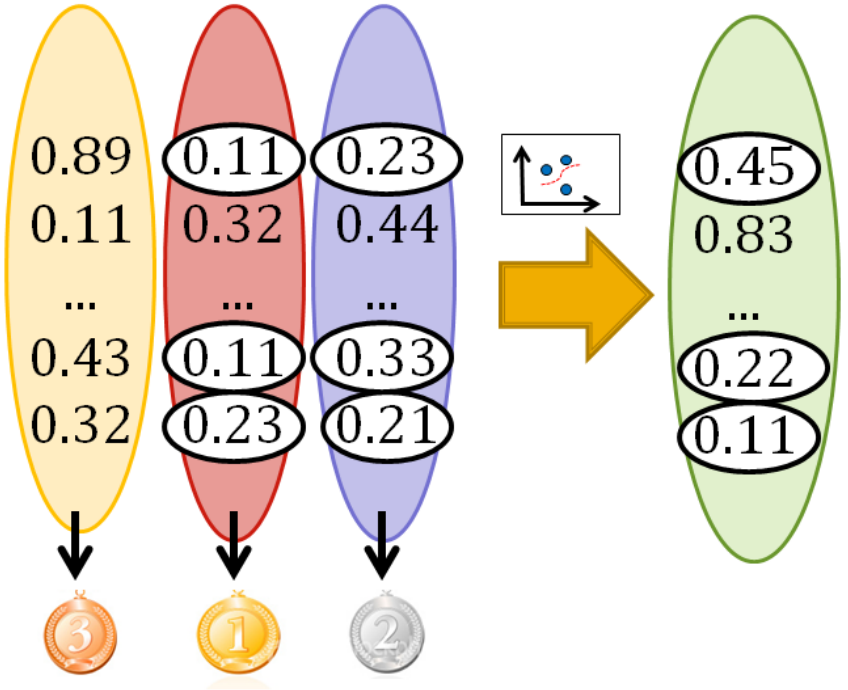
Use other algorithms in an ensemble setting

# Ensemble settings

In each regression problem, take a sample from both the available experiments and the considered transcription factors

Deduce feature importance scores for the predictors in the resulting regression problem

Repeat this step *X* times



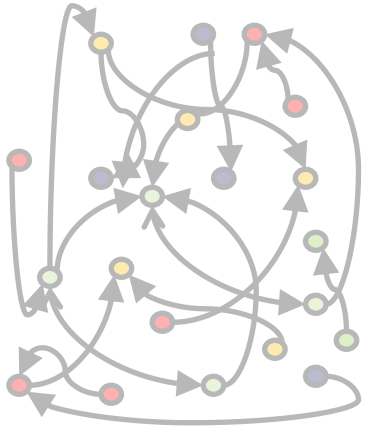
# Ensemble results

| = one shot<br>= ensemble<br>= inherent | DREAM4 | | DREAM5 artificial | | DREAM5 *E. coli.* | |
|---|---|---|---|---|---|---|
| | AUROC | AUPR | AUROC | AUPR | AUROC | AUPR |
| Elastic Net | 0.64 | 0.15 | -- | -- | -- | -- |
| Elastic Net | 0.73 | 0.19 | 0.78 | 0.28 | 0.63 | 0.11 |
| Lin-SVM | 0.57 | 0.03 | -- | -- | -- | -- |
| Lin-SVM | 0.78 | 0.17 | 0.78 | 0.23 | 0.61 | 0.12 |
| GENIE3 (RF) | 0.75 | 0.20 | 0.81 | 0.38 | 0.62 | 0.10 |
| Symbolic Regr. | 0.72 | 0.19 | 0.75 | 0.27 | 0.58 | 0.05 |

# Merging results

| = one shot<br>= ensemble<br>= inherent | DREAM4 | | DREAM5 artificial | | DREAM5 *E. coli.* | |
|---|---|---|---|---|---|---|
| | AUROC | AUPR | AUROC | AUPR | AUROC | AUPR |
| Elastic Net | 0.64 | 0.15 | -- | -- | -- | -- |
| Elastic Net | 0.73 | 0.19 | 0.78 | 0.28 | 0.63 | 0.11 |
| Lin-SVM | 0.57 | 0.03 | -- | -- | -- | -- |
| Lin-SVM | 0.78 | 0.17 | 0.78 | 0.23 | 0.61 | 0.12 |
| GENIE3 (RF) | 0.75 | 0.20 | 0.81 | 0.38 | 0.62 | 0.10 |
| Symbolic Regr. | 0.72 | 0.19 | 0.75 | 0.27 | 0.58 | 0.05 |
| Merge-avg. rank | 0.78 | 0.21 | 0.81 | 0.34 | 0.62 | 0.11 |

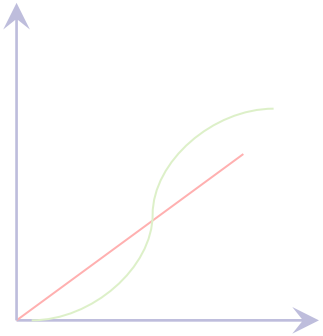
# Open issues



Biological problem

| | $g_1$ | ... | $g_{n-1}$ | $g_n$ |
|---|---|---|---|---|
| $s_1$ | 0.89 | ... | 0.11 | 0.23 |
| ... | | | | |
| $s_{p-1}$ | 0.43 | ... | 0.11 | 0.33 |
| $s_p$ | 0.32 | ... | 0.23 | 0.21 |

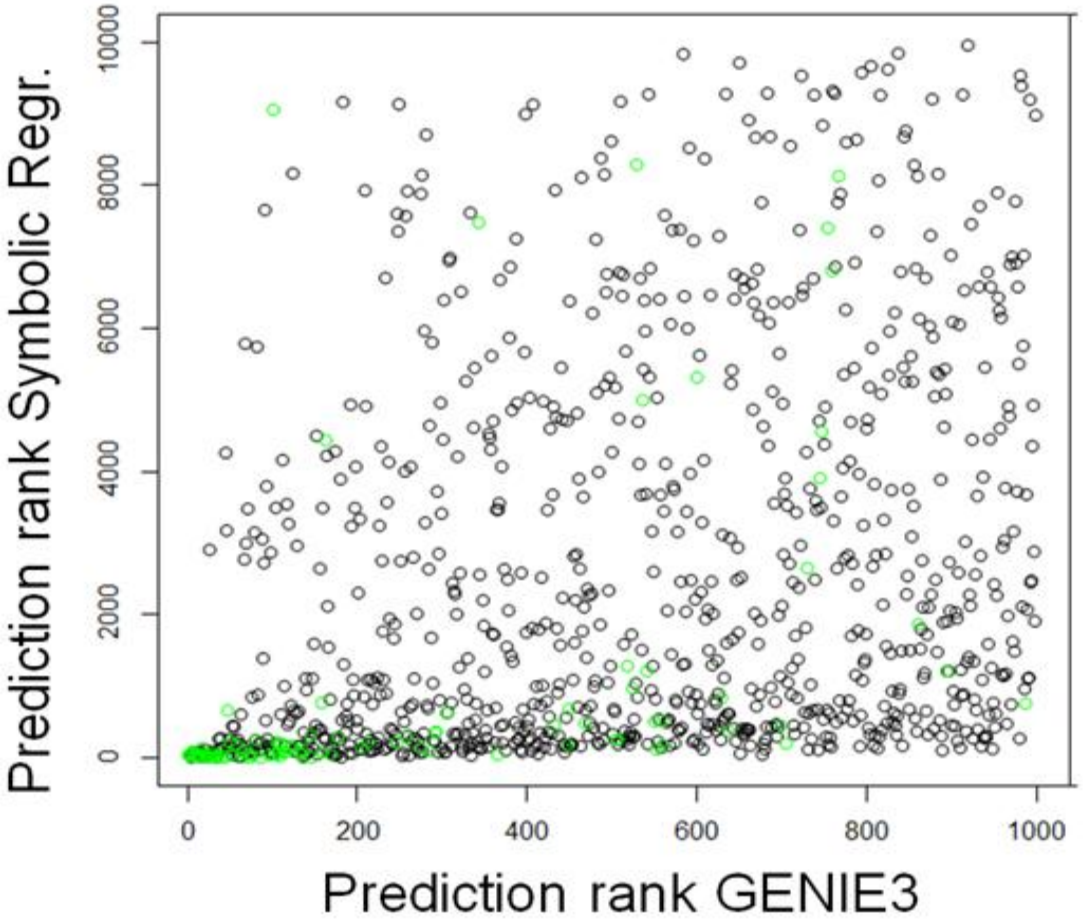
Machine learning problem



Results

**Open issues**

# Potential for improving predictions by merging

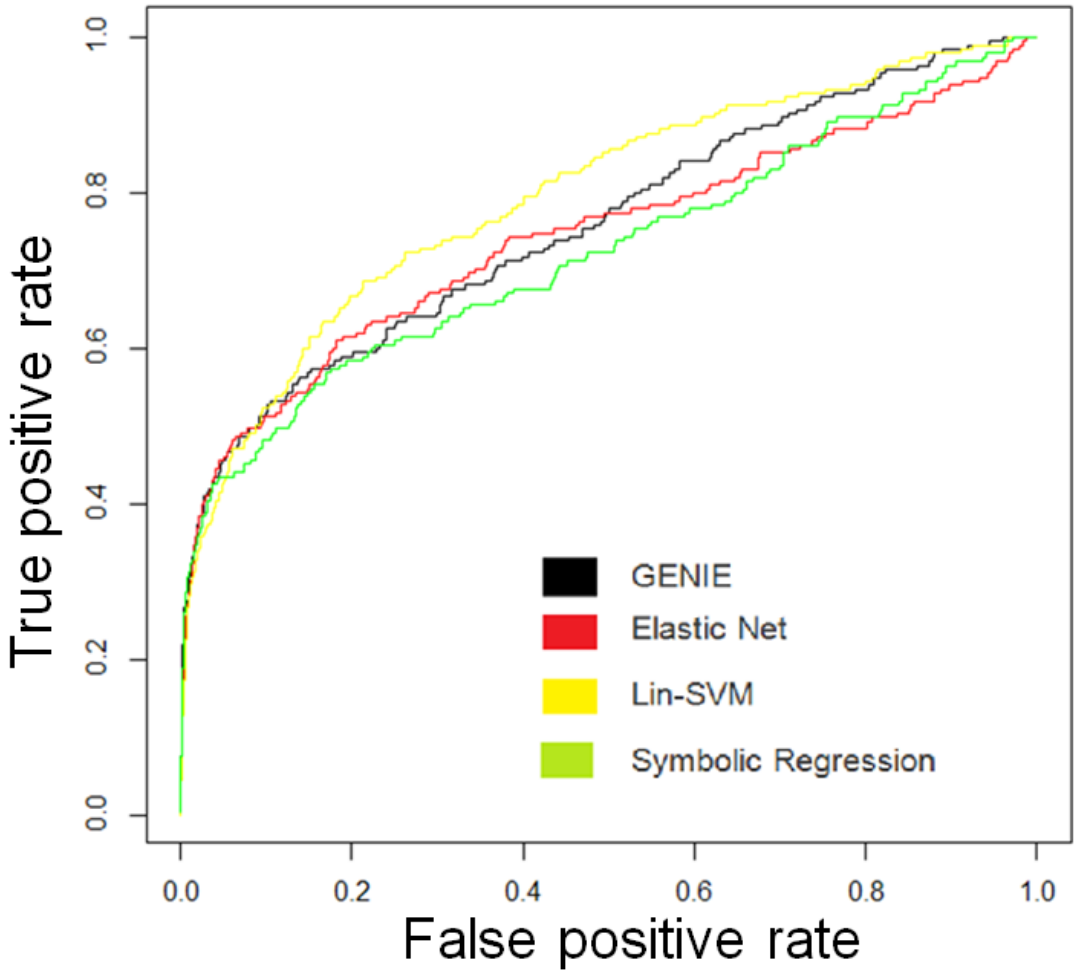



= true positive

= false positive

**Large rank differences between methods**

# Potential for improving predictions by merging




**ROC and PR curves sometimes indicate different algorithm characteristics**

# Current work- open issues

**Investigate bias of different variants:**

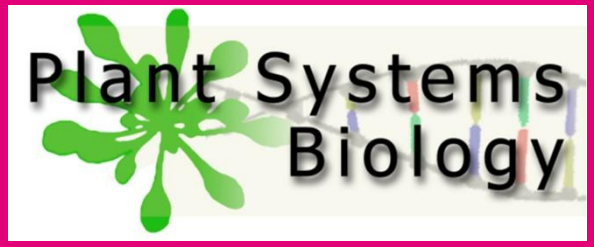
Fan-in , fan-out

Directionality

Edge of network vs. core of network

**Combine variants in one robust, better performing technique**




joeri.ruyssinck@intec.ugent.be
www.ibcn.intec.ugent.be
Internet Based Communication Networks and Services (IBCN)
Department of Information Technology (INTEC)
Ghent University - IBBT

ibbt
connect.innovate.create

UNIVERSITEIT GENT