



# Ensemble methods for bacterial network inference

Riet De Smet

## **Chairman**

Prof. Dr. Ir. J. Martens

## **Jury Members**

Prof. Dr. Ir. J. Vanderleyden

Dr. J. Ramon

Dr. T. Michoel (University of Freiburg)

Dr. K. McDowall (University of Leeds)

## **Promotors**

Prof. Dr. Ir. K. Marchal (promotor)

Prof. Dr. Ir. B. De Moor (co-promotor)



# Introduction



Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

## Ensemble methods for bacterial network inference

# Introduction

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

## Ensemble methods for bacterial network inference

*Goal*

# Introduction

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

## Ensemble methods for bacterial network inference

*Means*

*Goal*

# Introduction

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

## Ensemble methods for bacterial network inference

*Goal*

# *Escherichia coli*

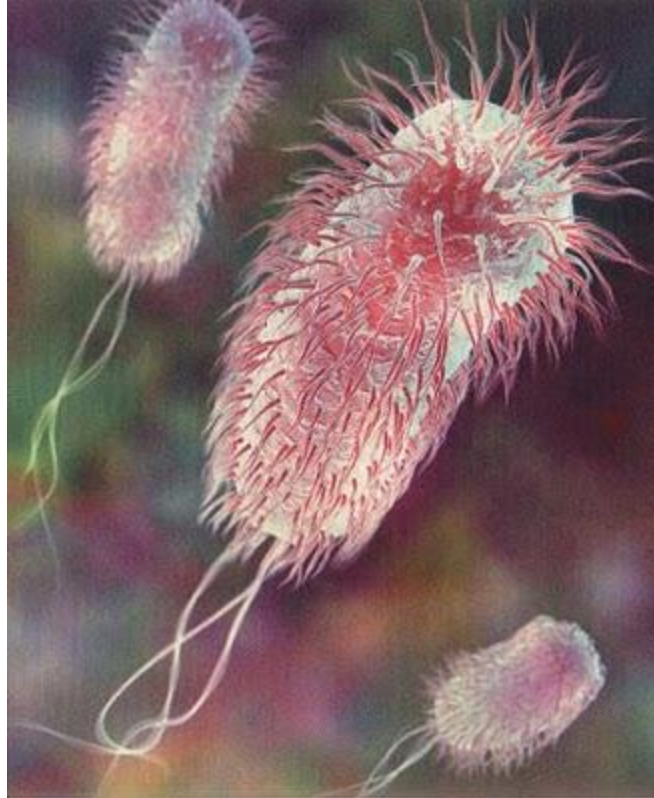
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# *Escherichia coli*

Introduction

Ensemble  
methods

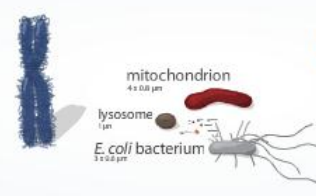
Module  
detection

Inference TRN

Conclusion

Small but important

X chromosome  
7  $\mu\text{m}$



baker's yeast  
3 x 4  $\mu\text{m}$

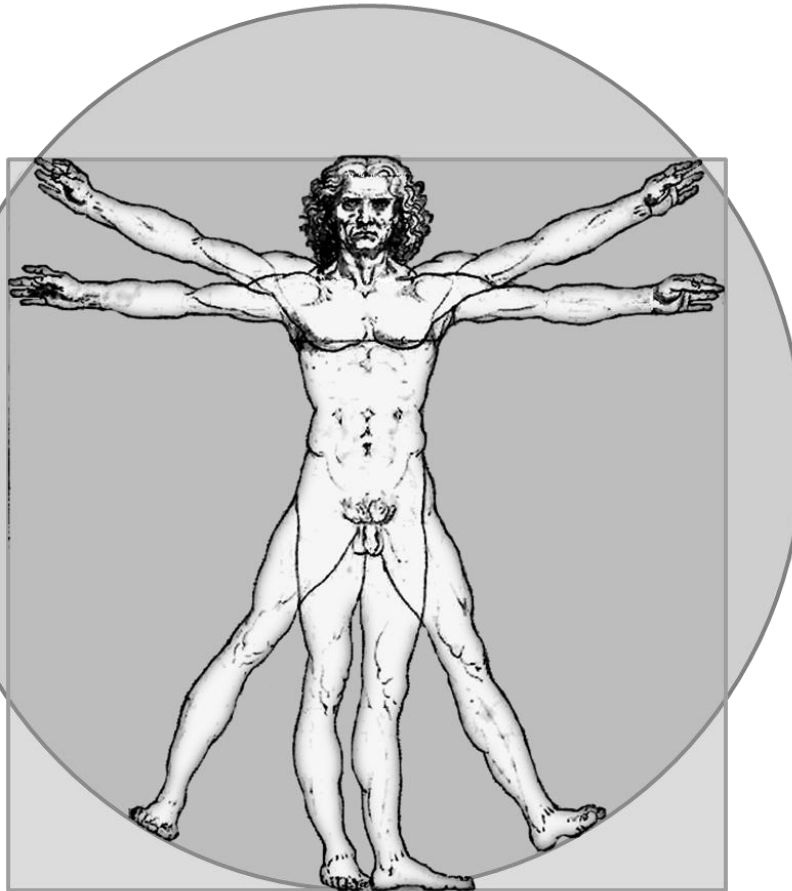


red blood cell  
8  $\mu\text{m}$

Genetic Science Learning Center Utah, University of Utah

# *Escherichia coli*

In sickness and in health



Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# *Escherichia coli*

In sickness and in health

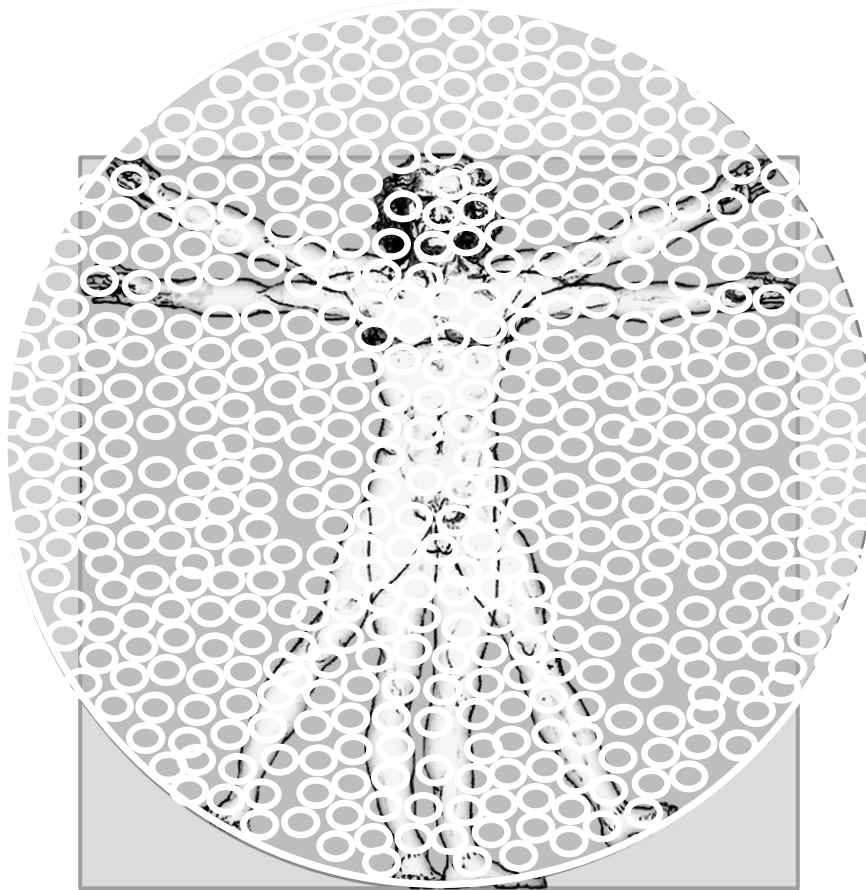
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# Escherichia coli

In sickness and in health

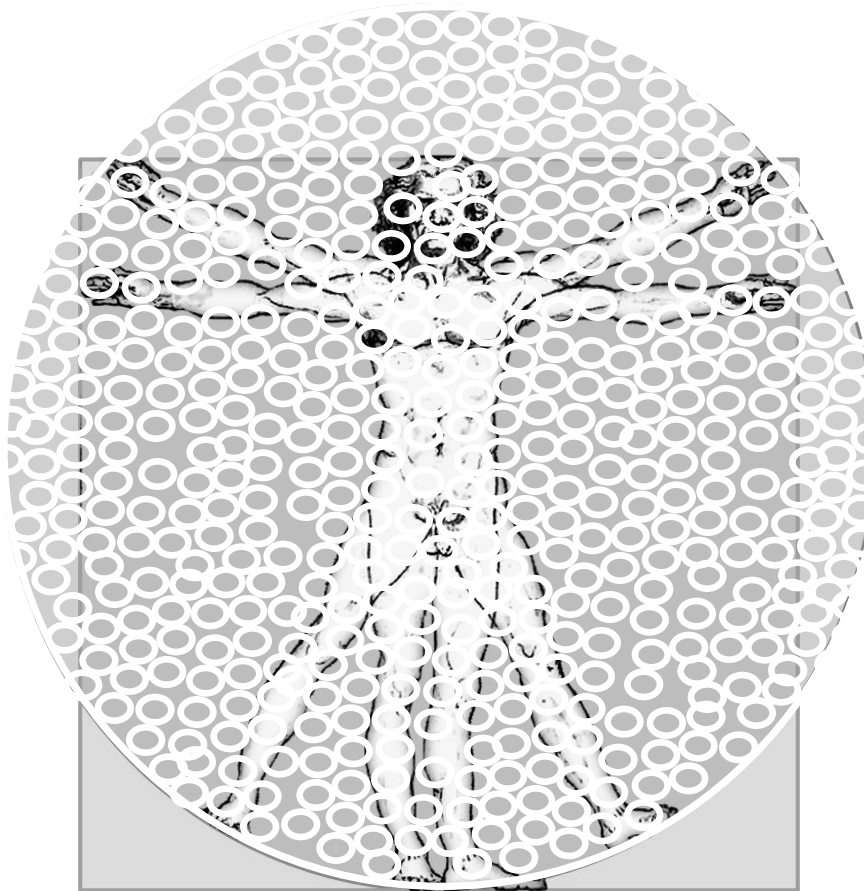
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



INTERNATIONAL FOOD SAFETY NETWORK INFOSHEET SEPT 12/08  
www.foodsafety.ksu.edu foodsafetyinfosheets.ksu.edu



- E. coli O111 is usually found in human and animal poop.
- Keep the poop out of the food you serve by practicing good hygiene and handwashing.
- E. coli can be passed on by food handlers who don't have symptoms.
- E. coli outbreaks have also been linked to fresh produce and other ready-to-eat foods.

An outbreak in Locust Grove, Oklahoma featuring an uncommon strain of E. coli has been linked to at least 248 people. Sixty-four people have been hospitalized and a 26 year-old gospel singer, Chad Ingle, has died. Sixteen of those hospitalized, including nine children, have received dialysis treatment.

The state has announced that it has not found E. coli O111 on surfaces or in leftovers at the restaurant. Many of the victims ate at Country Cottage between August 15 and 17. Ten attendees of a church tea catered by Country Cottage have also become ill with E. coli O111.

State Epidemiologist Kristy Bradley said that the lack of a specific food points to the possibility that there was an E. coli-infected food handler at the root of the outbreak. The restaurant is a large buffet-style operation with 58 food handlers. While most employees have a primary station, it is not unusual to handle a number of foods.

Several victims of this outbreak and their families have contacted attorneys looking to recoup at least the cost of their medical bills.

The E. coli O111 outbreak is believed to be the largest of its kind in



# *Escherichia coli*

Jack of all trades

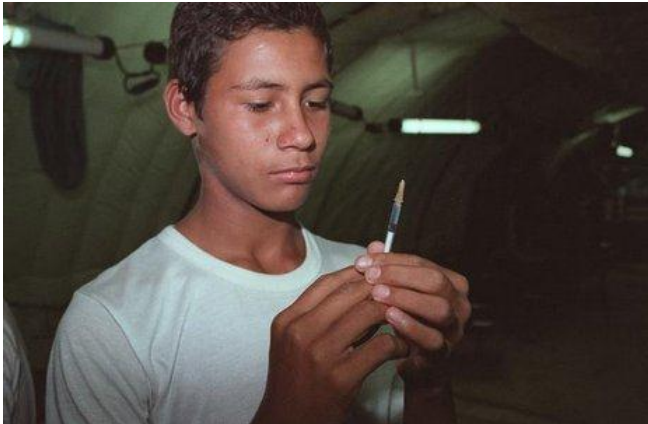
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



Insulin production

# Escherichia coli

Introduction

Jack of all trades

Ensemble methods

Module detection

Inference TRN

Conclusion



Insulin production

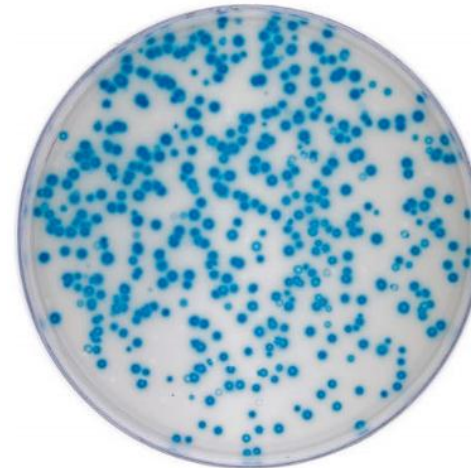
## DeMorgen.be

NIEUWS | SPORT | GELD | MUZIEK | DE GEDACHTE | PLANET WATCH | **TECHNOLOGY** | MAGAZINE | CULTUUR & MEDIA

INTERNET | GAMES | MULTIMEDIA

## Studenten stoppen 900 TB data in gram bacteriën

29/11/10, 17u13



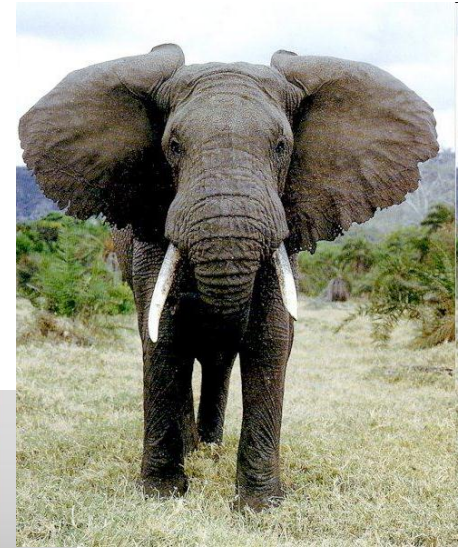
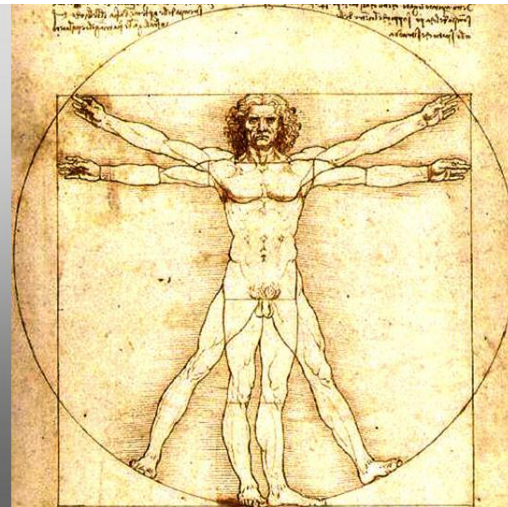
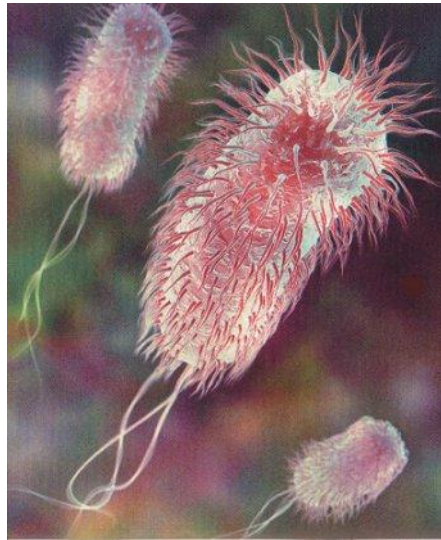
Wetenschappers hebben een biologisch opslagsysteem ontwikkeld dat enorme hoeveelheden data kan bevatten. Met het onderzoek behaalde het team de eerste plaats op de International Genetically Engineered Machine (iGEM)-wedstrijd, georganiseerd door het Massachusetts Institute of Technology.

Is *E. coli* our future hard disk?



# Escherichia coli

“Ce qui est vrai pour le colibacille est vrai pour l’éléphant”  
*J. Monod (Nobel laureate 1965)*



Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

# Transcriptional regulatory network

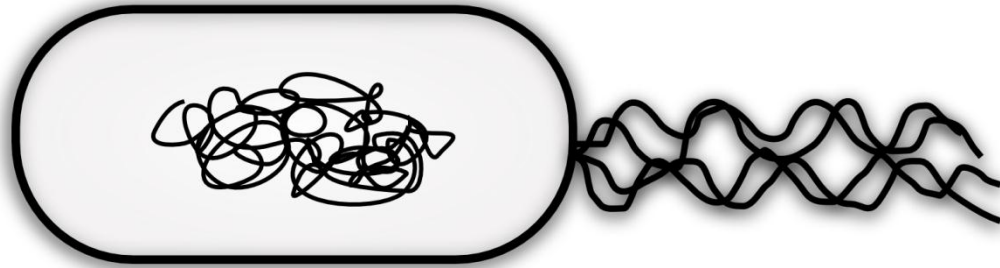
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# Transcriptional regulatory network

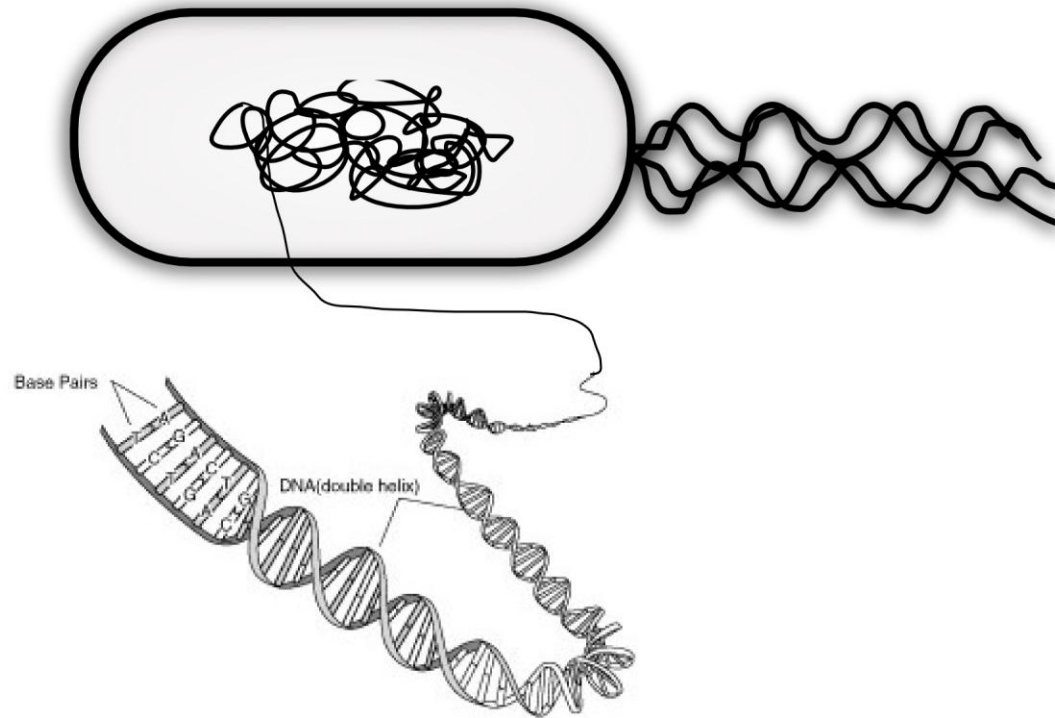
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# Transcriptional regulatory network

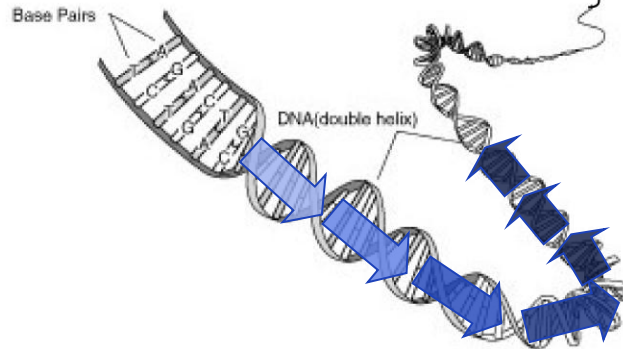
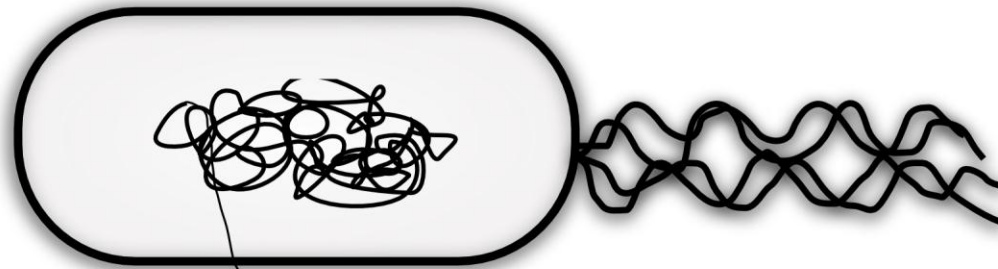
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



*E. coli* genome:

- 4,639,675 base pairs
  - 4494 genes
- (source NCBI)

---

## ARTICLE

### The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner,\* Guy Plunkett III,\* Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao

The 4,639,221-base pair sequence of *Escherichia coli* K-12 is presented. Of 4288 protein-coding genes annotated, 38 percent have no attributed function. Comparison with five other sequenced microbes reveals ubiquitous as well as narrowly distributed gene families; many families of similar genes within *E. coli* are also evident. The largest

The first 1.92 Mb (13, 14), positions 2,686,777 to 4,639,221 [in base pairs (bp)], was sequenced from our overlapping set of 15- to 20 kb MG1655 lambda clones (15)



# Transcriptional regulatory network

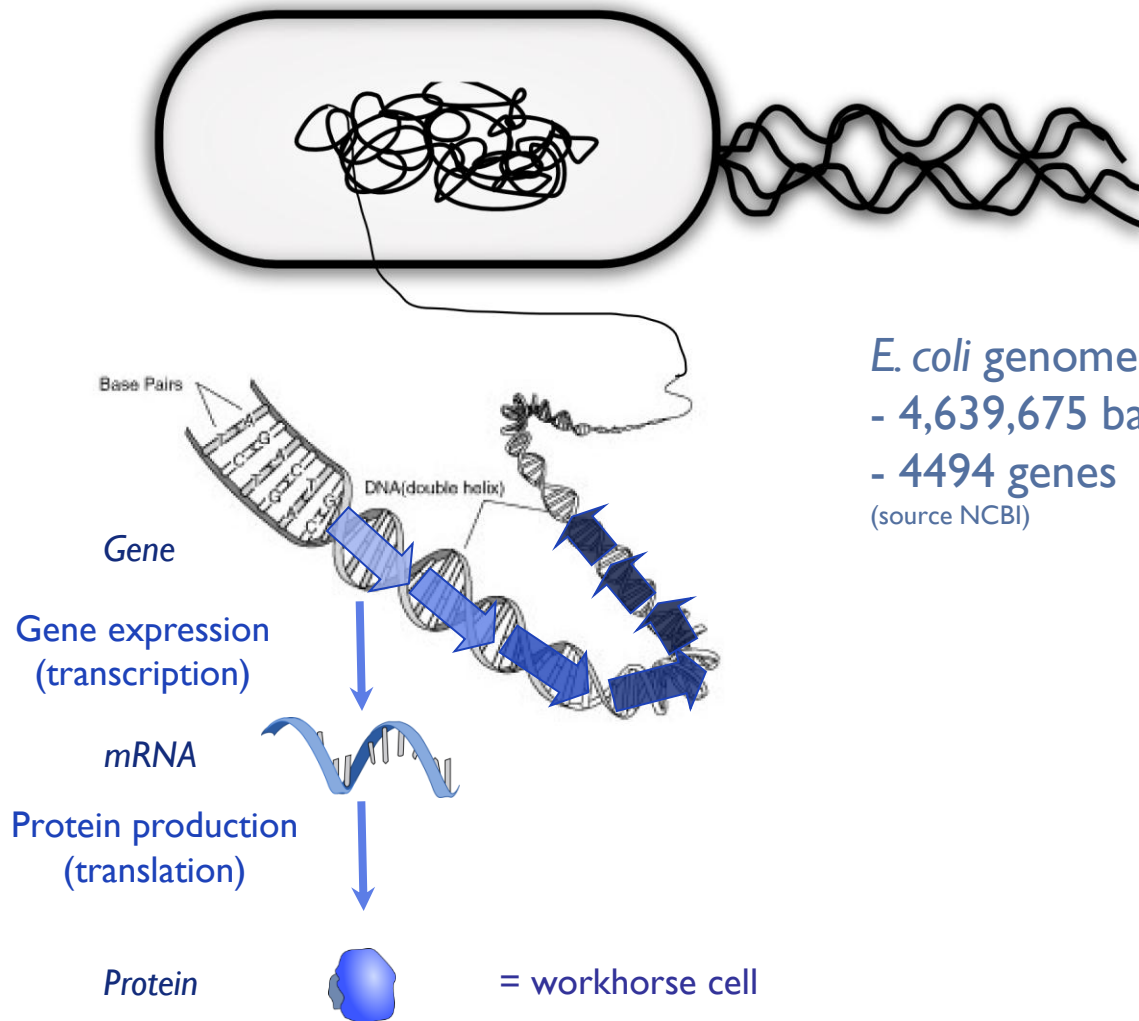
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



*E. coli* genome:

- 4,639,675 base pairs
  - 4494 genes
- (source NCBI)

# Transcriptional regulatory network

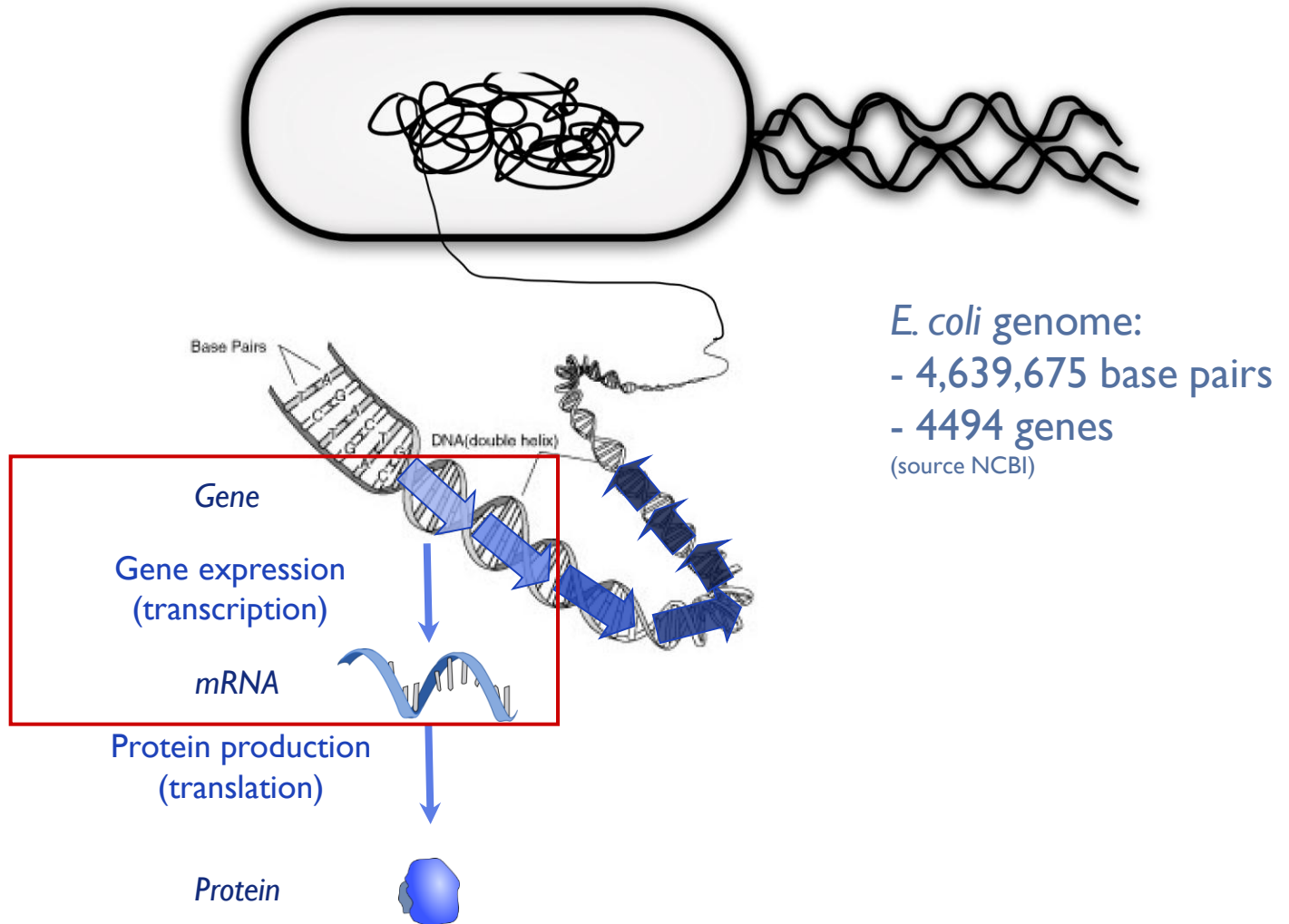
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Transcriptional regulatory network

Introduction

Ensemble methods

Module detection

Inference TRN

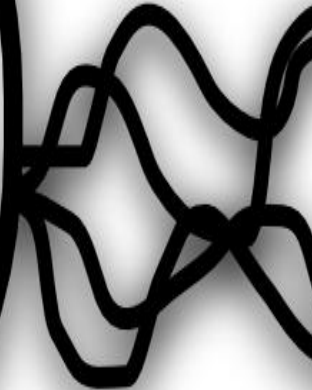
Conclusion

Needs sugar

*E. coli* genome:

- ~160 confirmed TFs
- ~300 estimated TFs

Transcription factor



# Transcriptional regulatory network

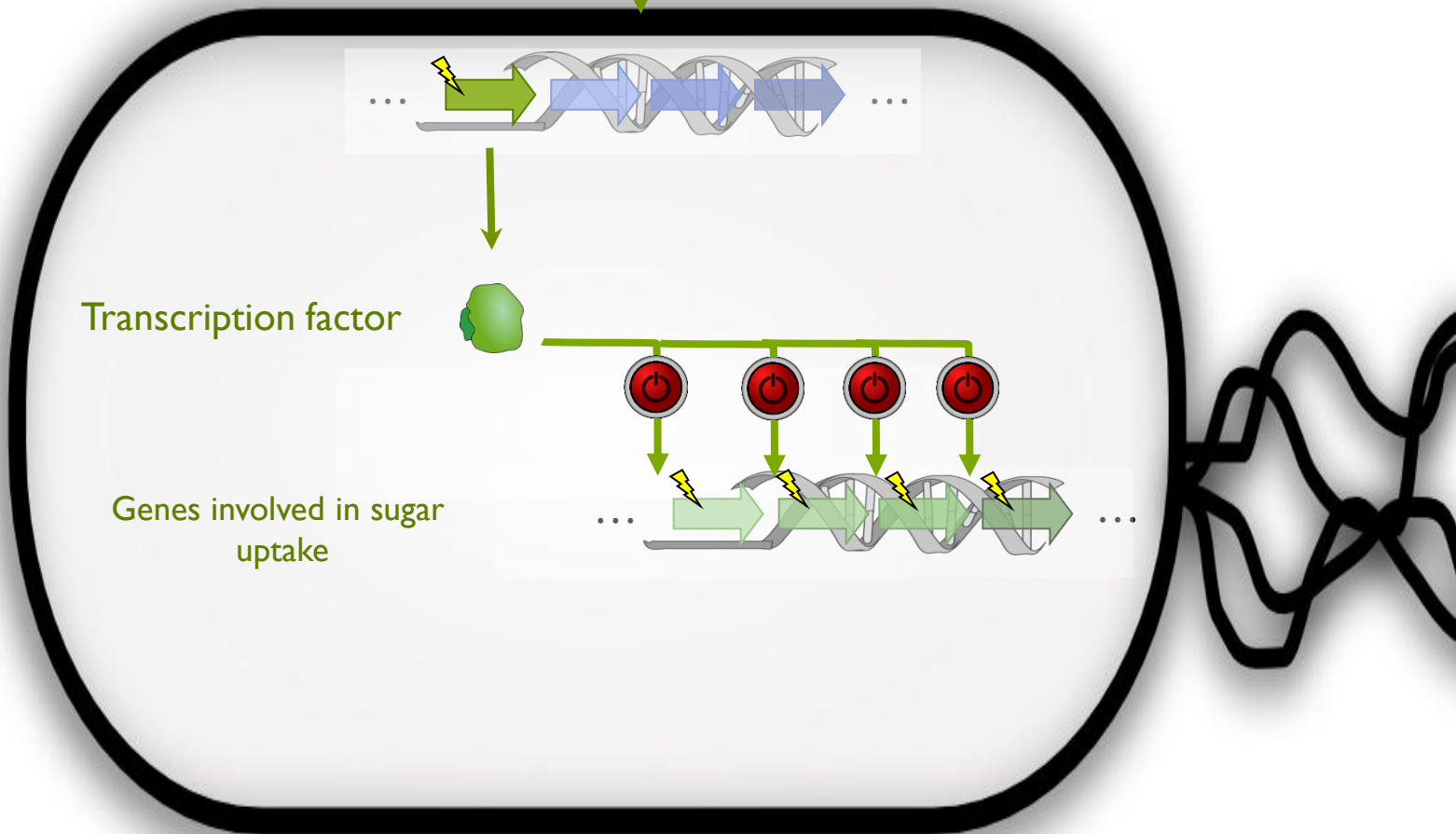
*E. coli* genome:

- ~160 confirmed TFs
- ~300 estimated TFs



Transcription factor

Genes involved in sugar uptake



Introduction

Ensemble methods

Module detection

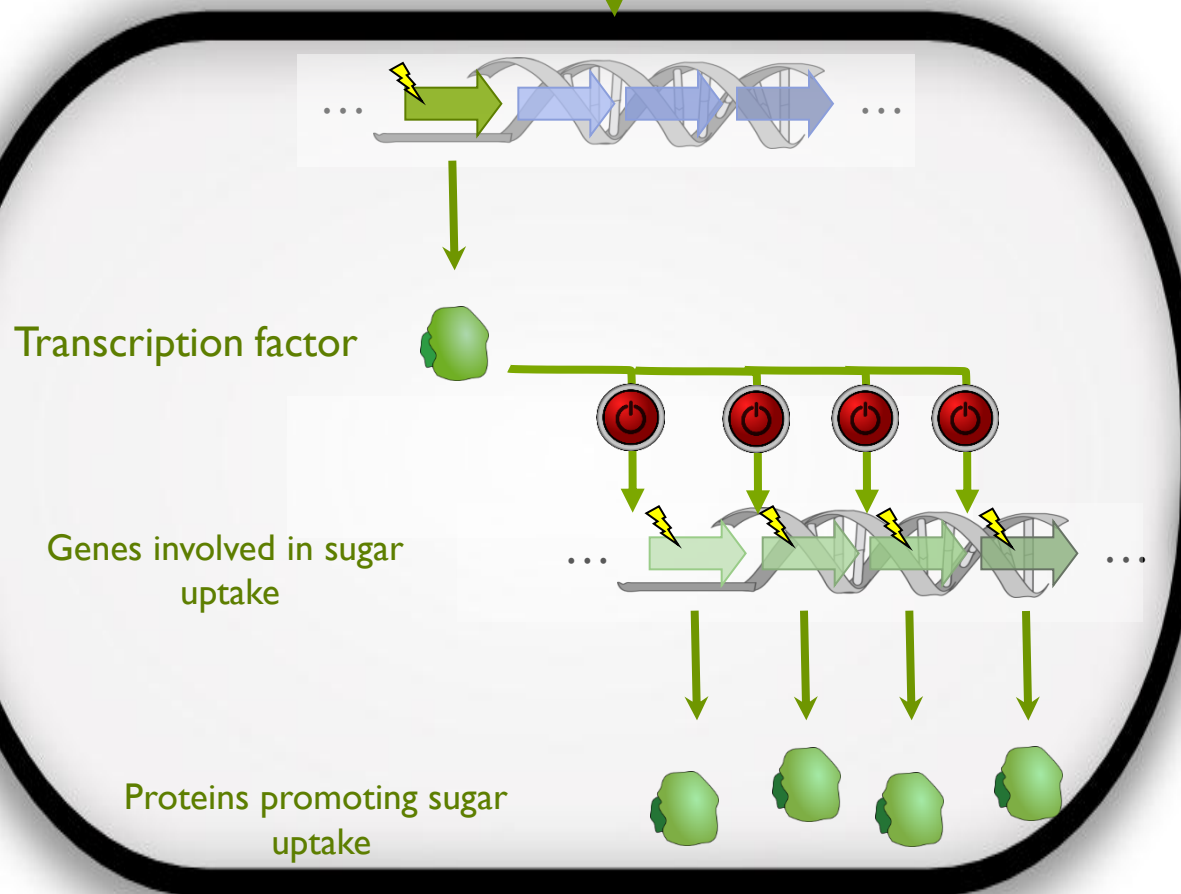
Inference TRN

Conclusion

# Transcriptional regulatory network

*E. coli* genome:

- ~165 confirmed TFs
- ~200 estimated TFs



Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Transcriptional regulatory network

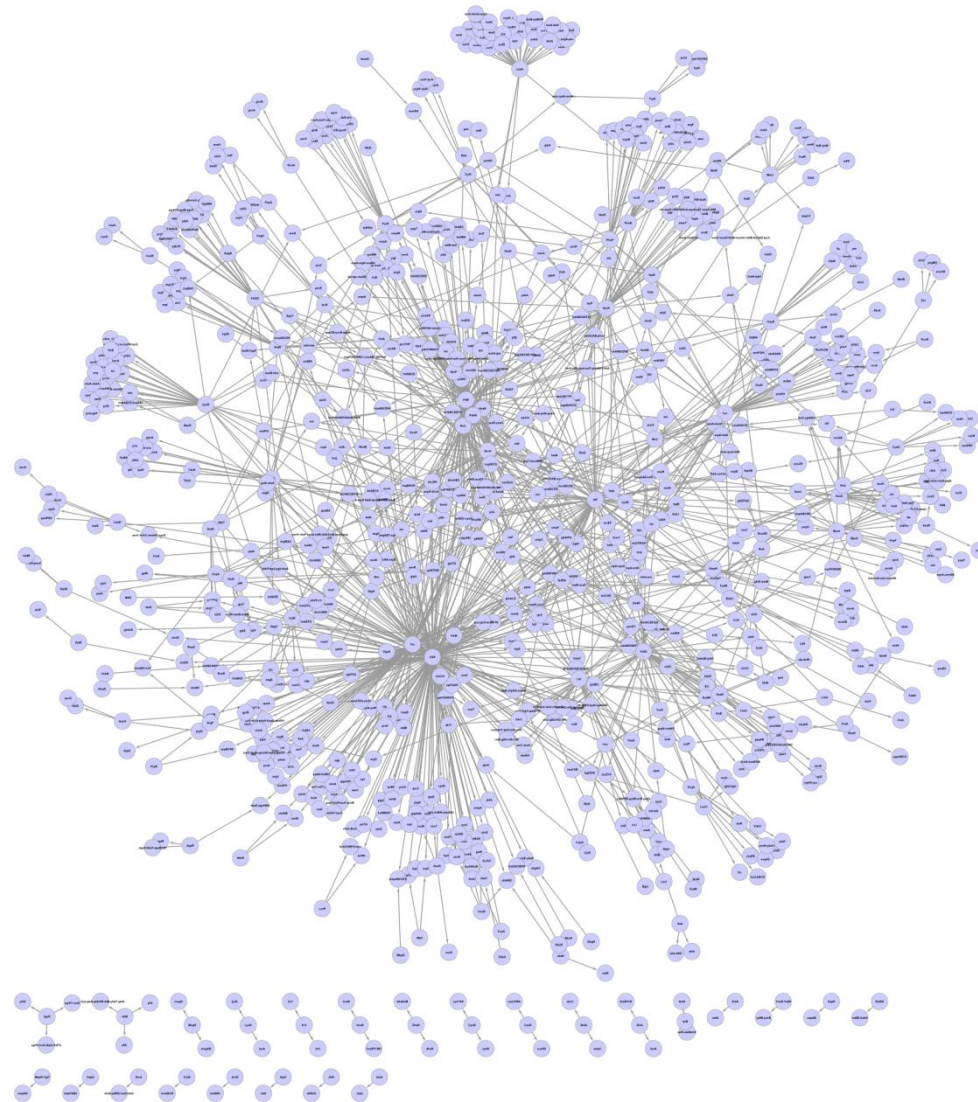
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

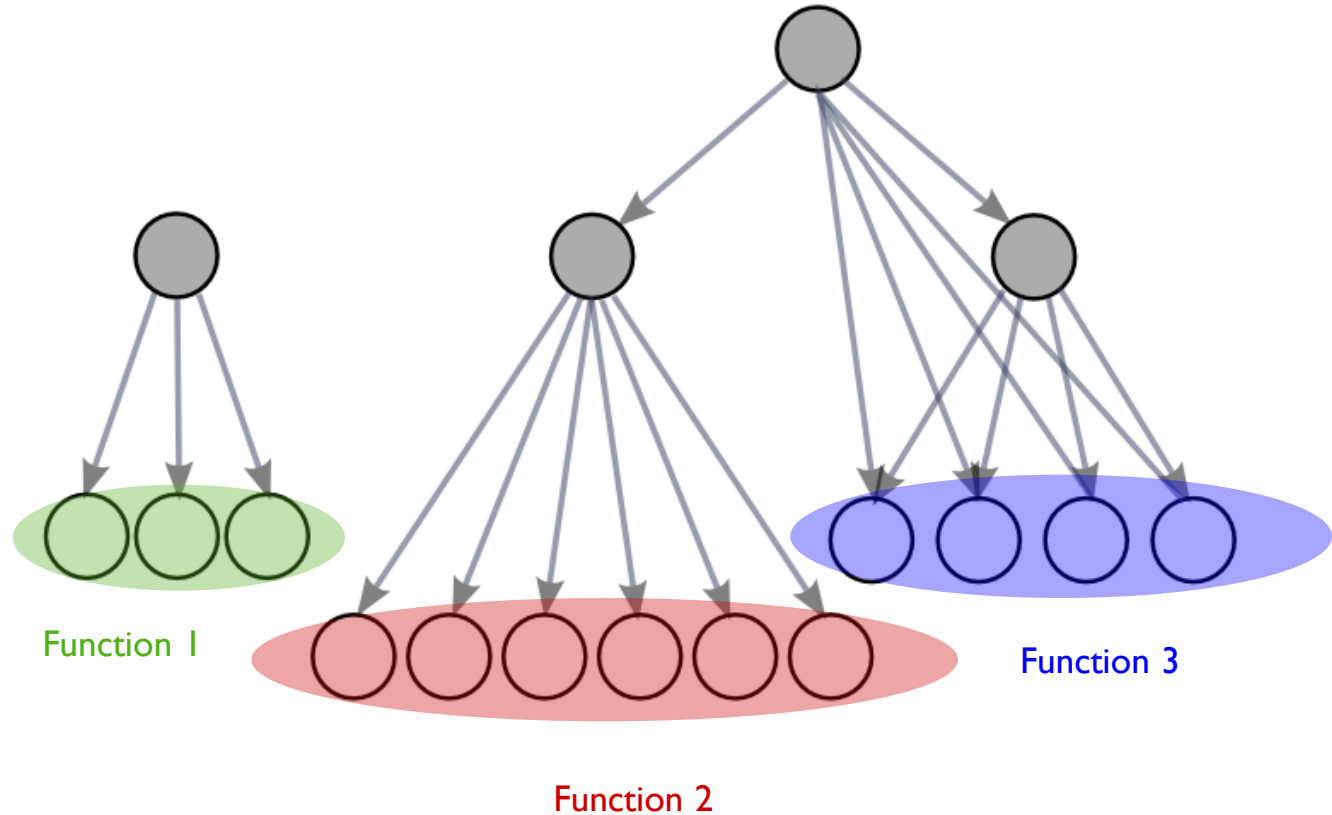
Conclusion





# Transcriptional regulatory network

Modular



Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Transcriptional regulatory network

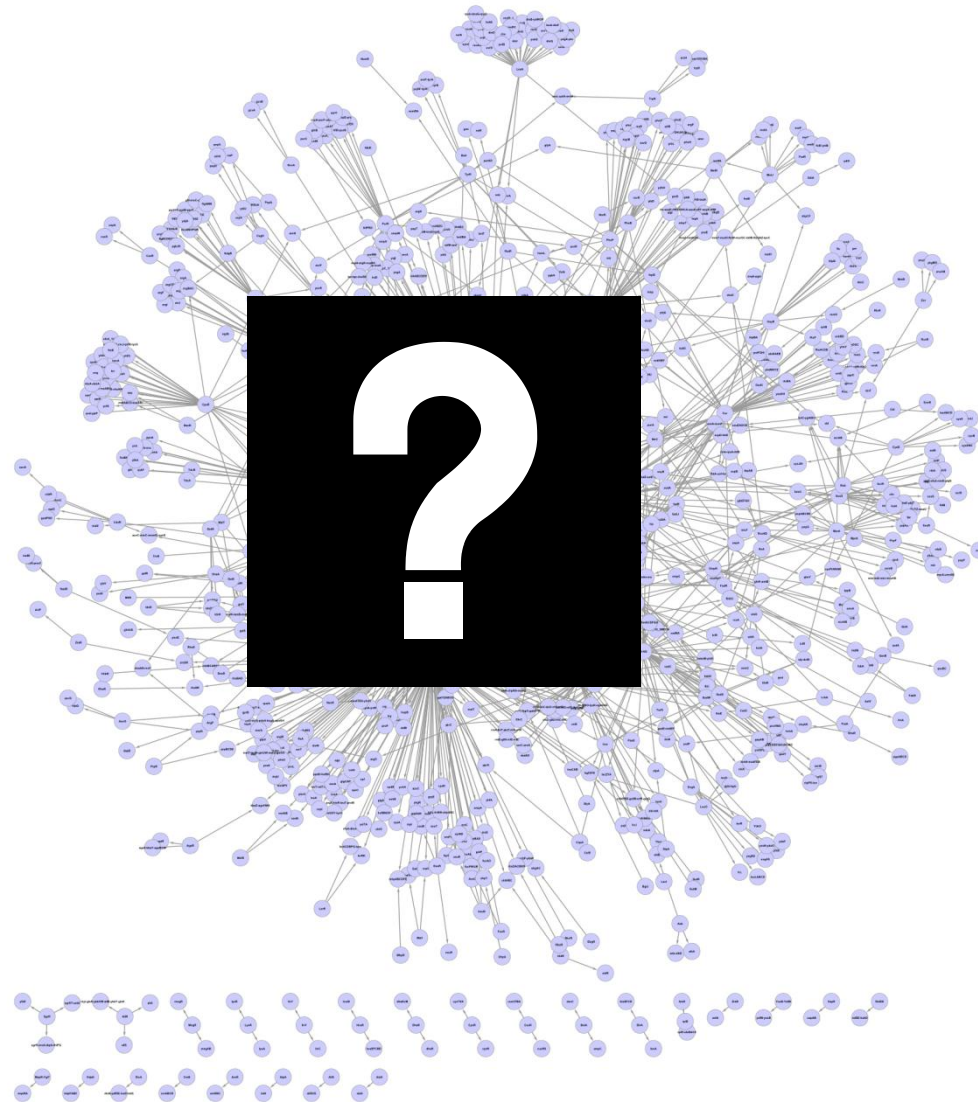
Introduction

Ensemble  
methods

Module  
detection

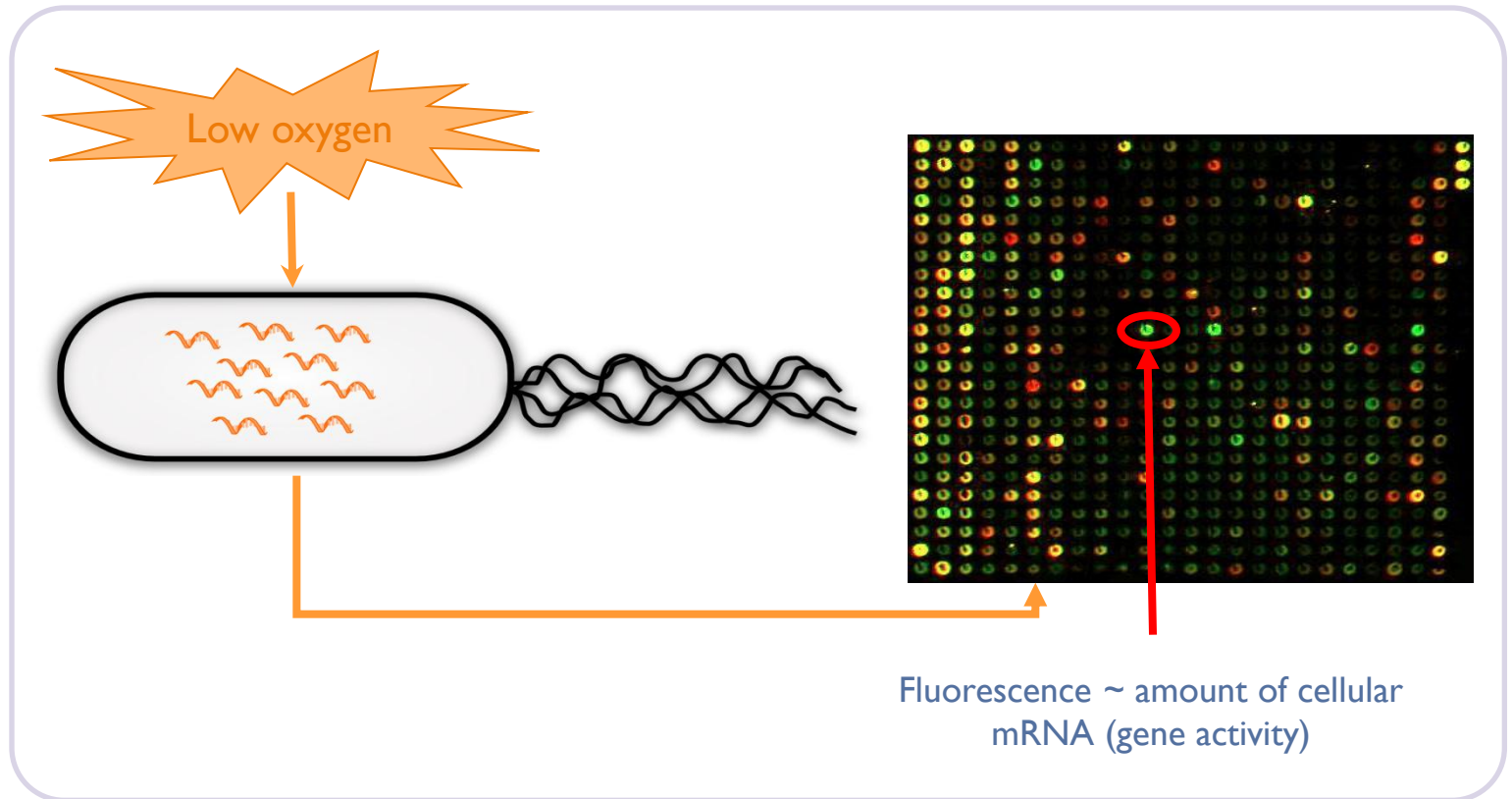
Inference TRN

Conclusion



# Microarray data

Snapshot of gene activity under a particular condition



Introduction

Ensemble methods

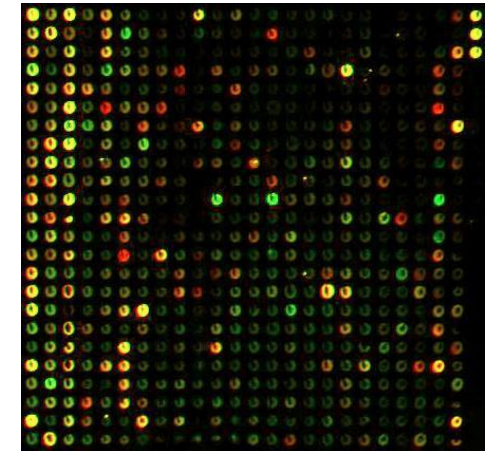
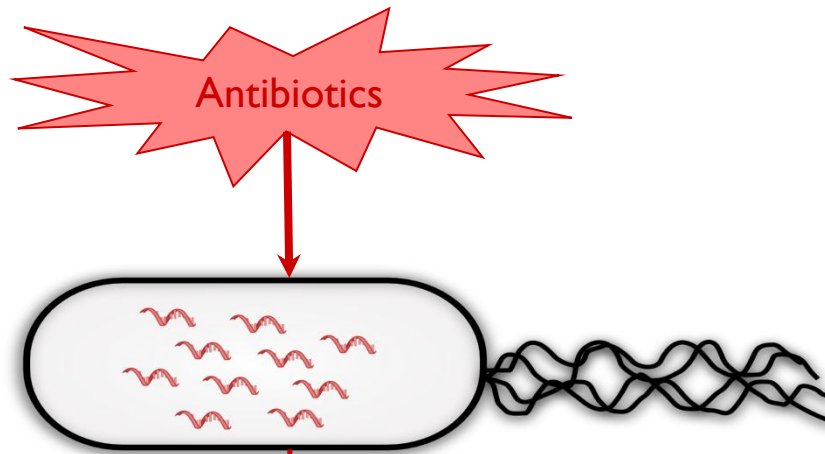
Module detection

Inference TRN

Conclusion

# Microarray data

Snapshot of gene activity under a particular condition



Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Microarray data

Snapshot of gene activity under a particular condition

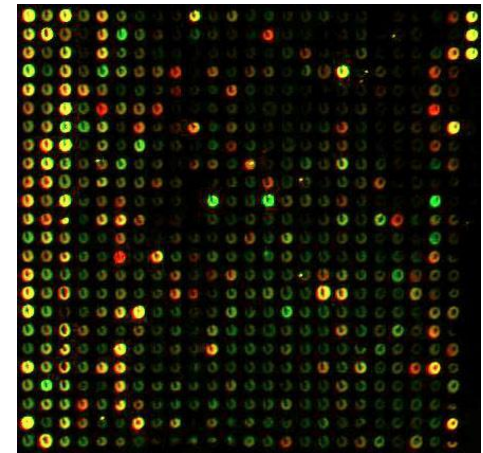
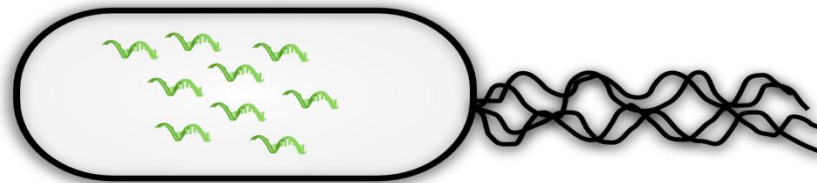
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion





# Microarray data

## Gene expression compendium

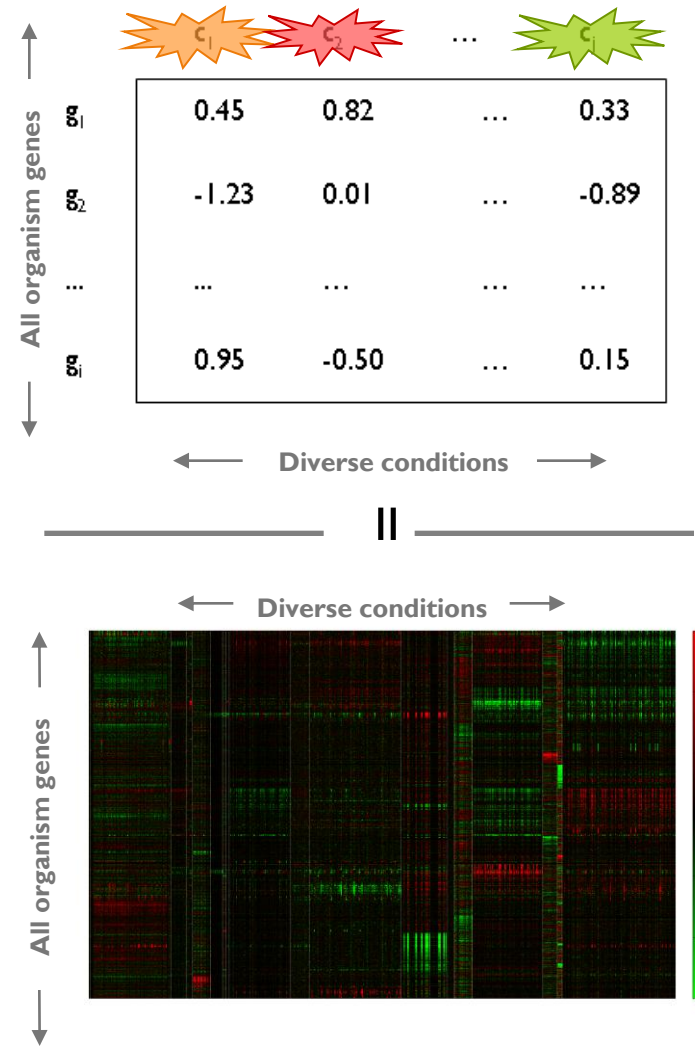
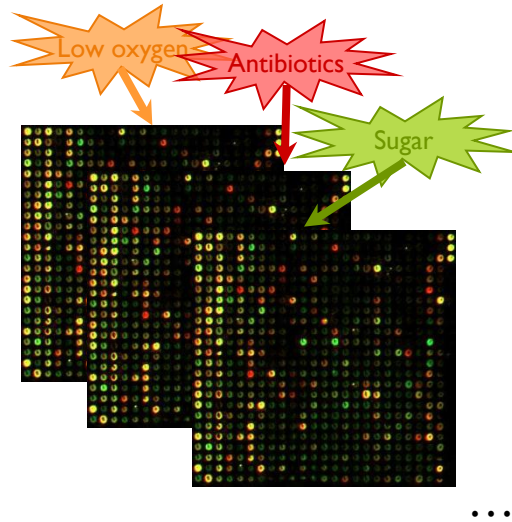
Introduction

Ensemble methods

Module detection

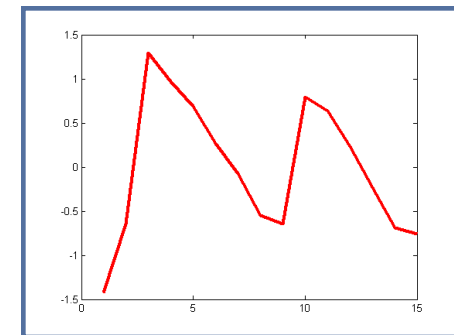
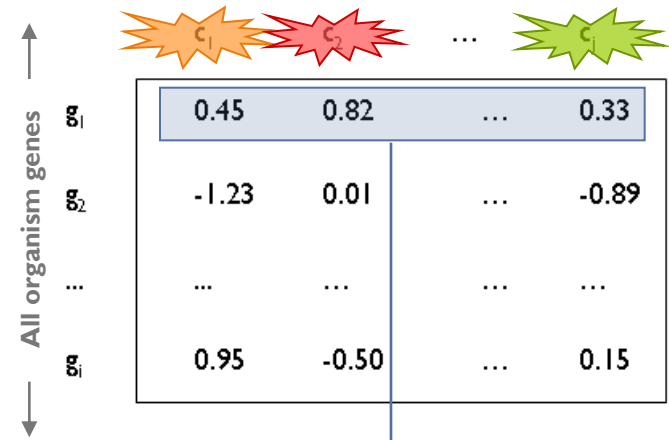
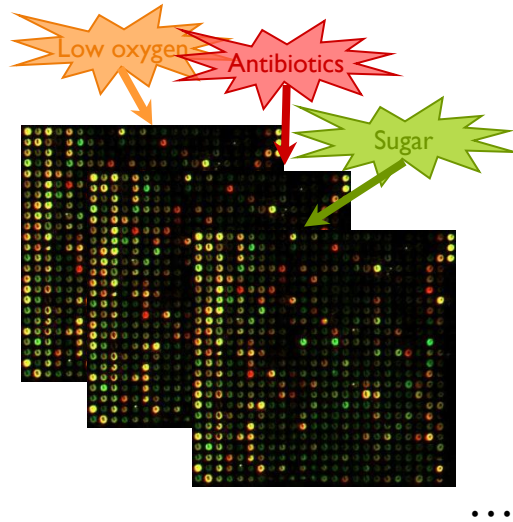
Inference TRN

Conclusion



# Microarray data

## Gene expression compendium



Gene expression profile

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Introduction

## Network inference

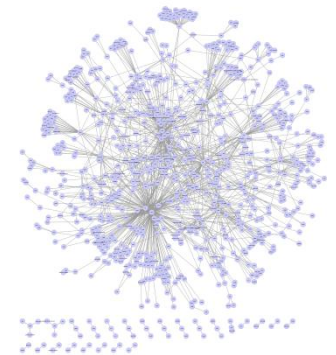
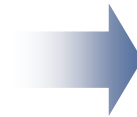
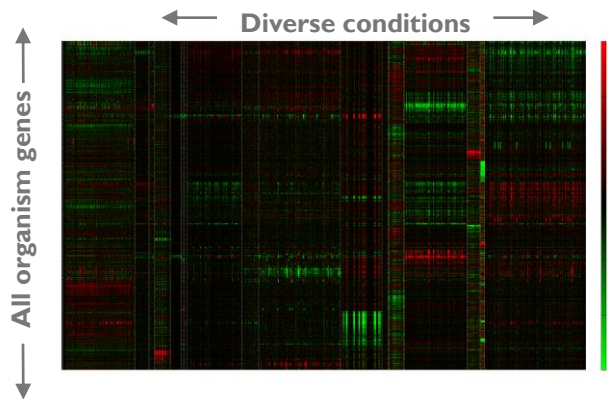
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# Introduction

## Network inference

Prediction of Transcriptional Regulatory Network (TRN) from  
microarray data

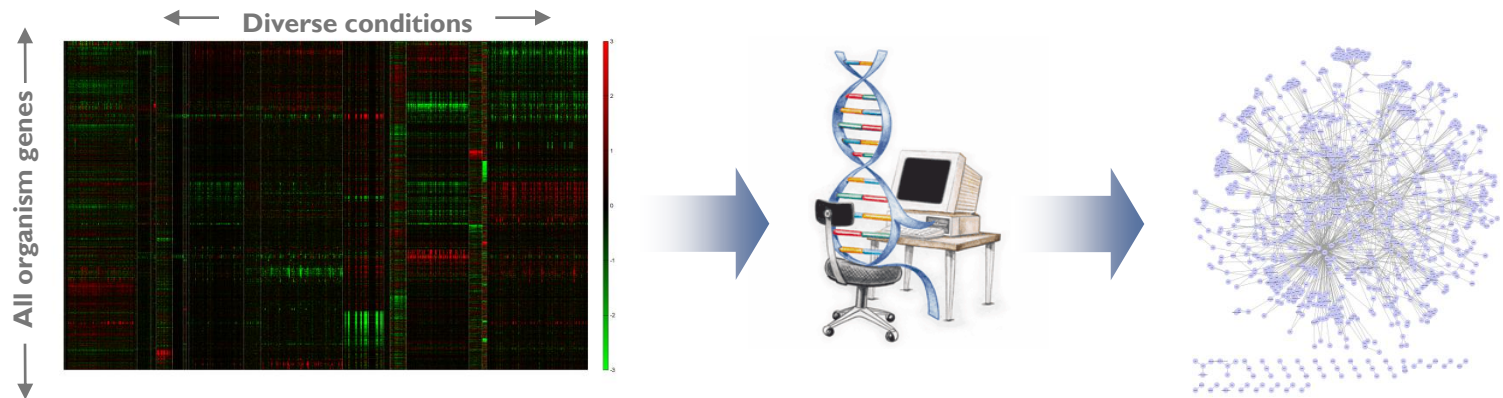
Introduction

Module  
detection

Inference  
TRN

Conclusion

Perspectives





# Network inference

## I. Module detection

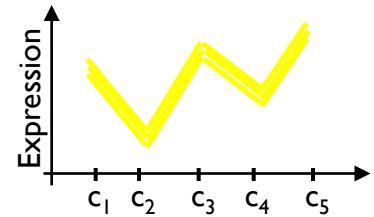
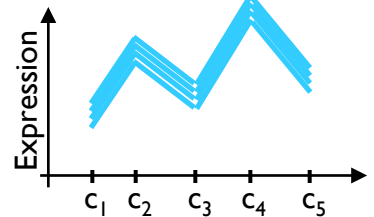
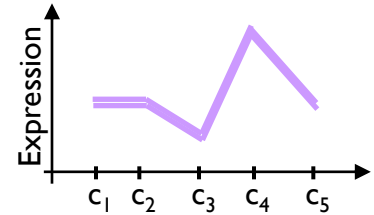
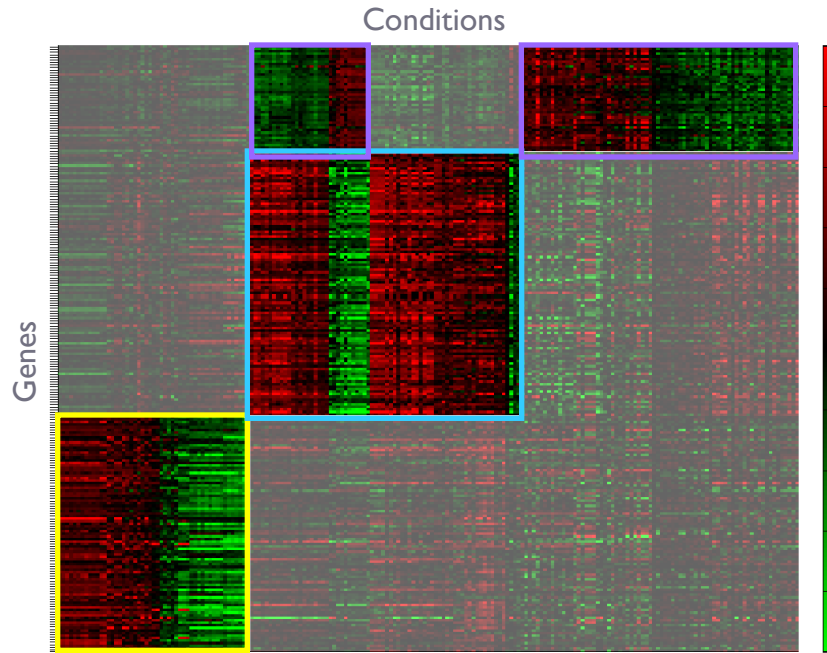
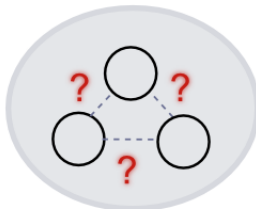
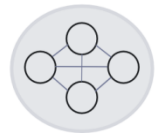
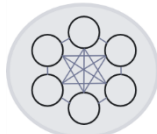
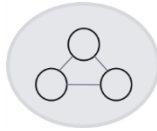
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

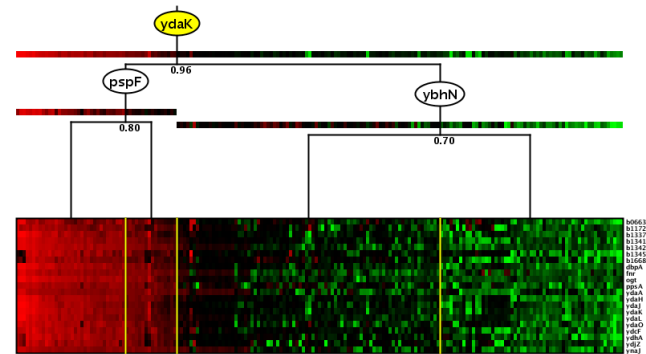
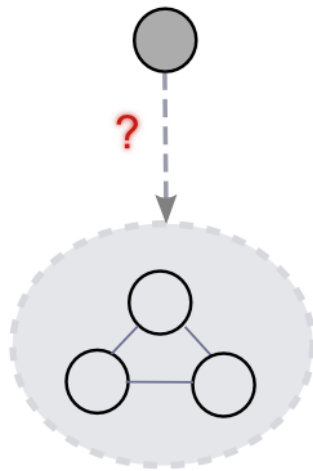


(Bi)clustering

— = coexpression

# Network inference

## II. Inference TRN



→ = regulation

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Introduction

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

## Ensemble methods for bacterial network inference

*Means*

# Ensemble methods

Introduction

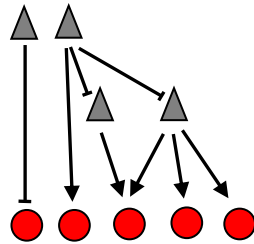
Ensemble methods

Module detection

Inference TRN

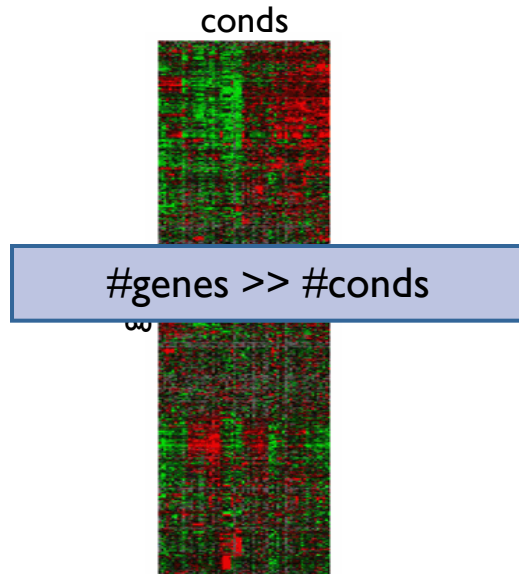
Conclusion

Biological complexity



Large search space

Limits data



Limited number of samples

# Ensemble methods

Introduction

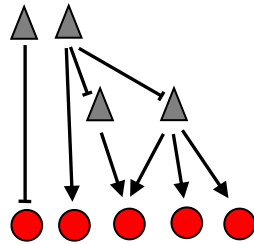
Ensemble methods

Module detection

Inference TRN

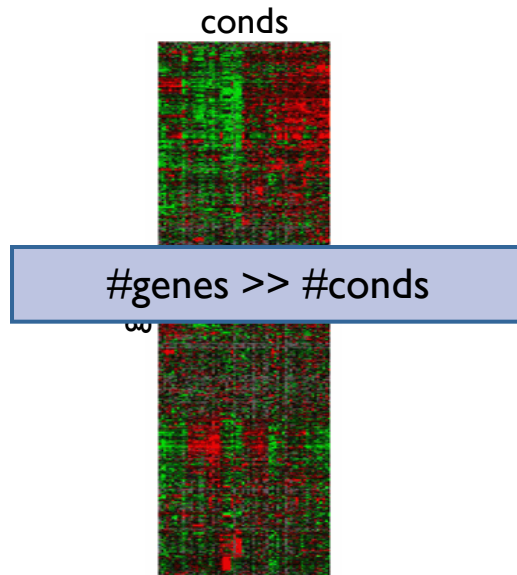
Conclusion

Biological complexity



Large search space

Limits data



System heavily **underdetermined**

Limited number of samples

# Ensemble methods

Many possible solutions exist that each explain the data equally well

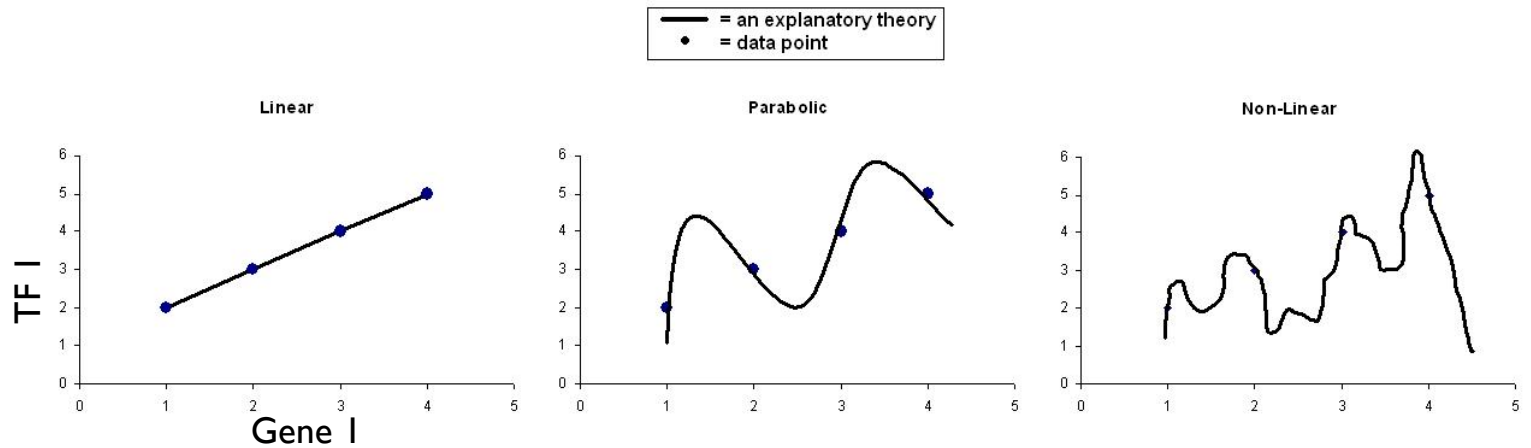
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Ensemble methods

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

Strategies to deal with problem of underdetermination

- Constrain search space: **Chapter 6**
- Consider ensemble of solutions: **Chapter 3, 5**

# Ensemble methods

Introduction

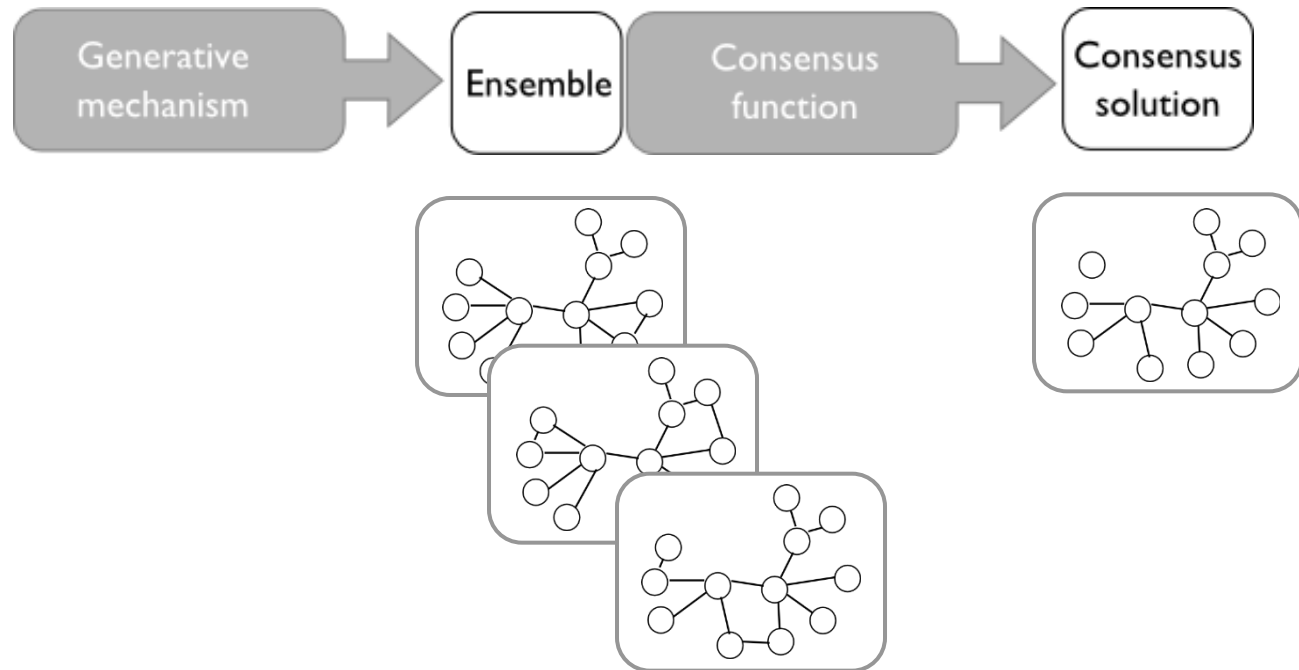
Ensemble methods

Module detection

Inference TRN

Conclusion

- Consider multiple possible predictions from the data (Generative mechanism)
- Combine predictions in an intelligent way (Consensus function)





# Ensemble methods

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

- Consider multiple possible predictions from the data (Generative mechanism)
  - existing methods (one or multiple methods)
- Combine predictions in an intelligent way (Consensus function)
  - Goal I: improve accuracy
  - Goal II: extend breadth predictions

Goal PhD: improve shortcomings existing network inference methods by using ensemble strategies

# Ensemble methods

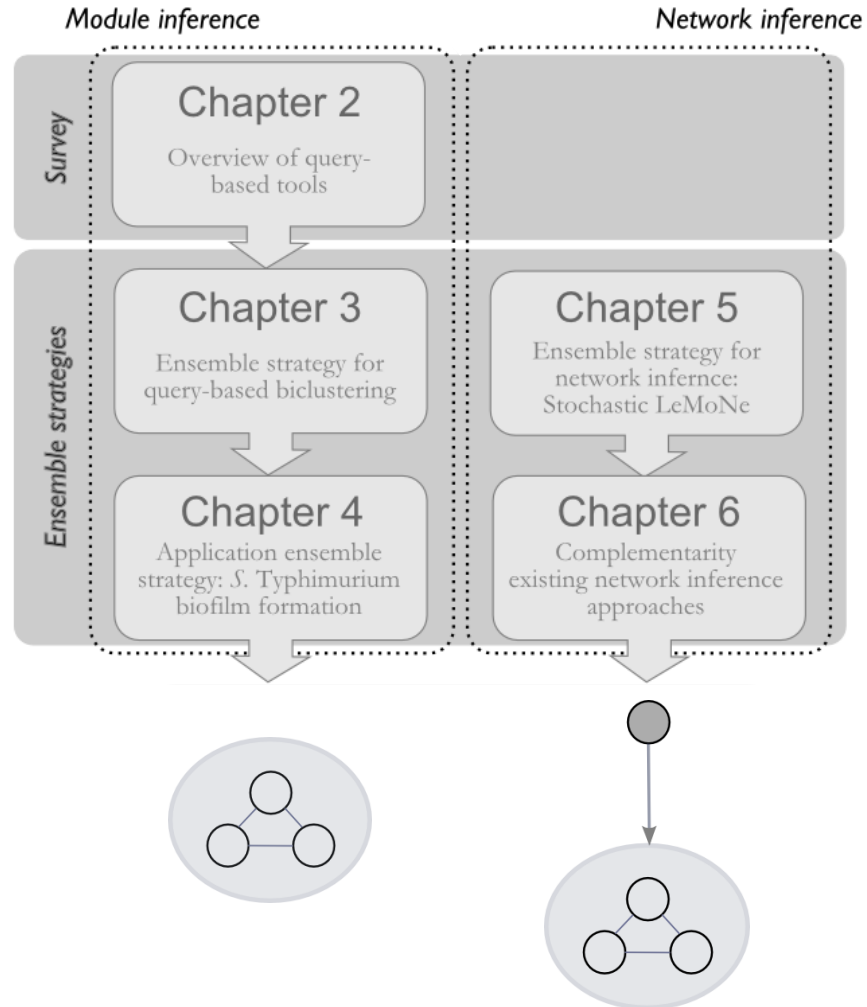
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Ensemble method for module detection

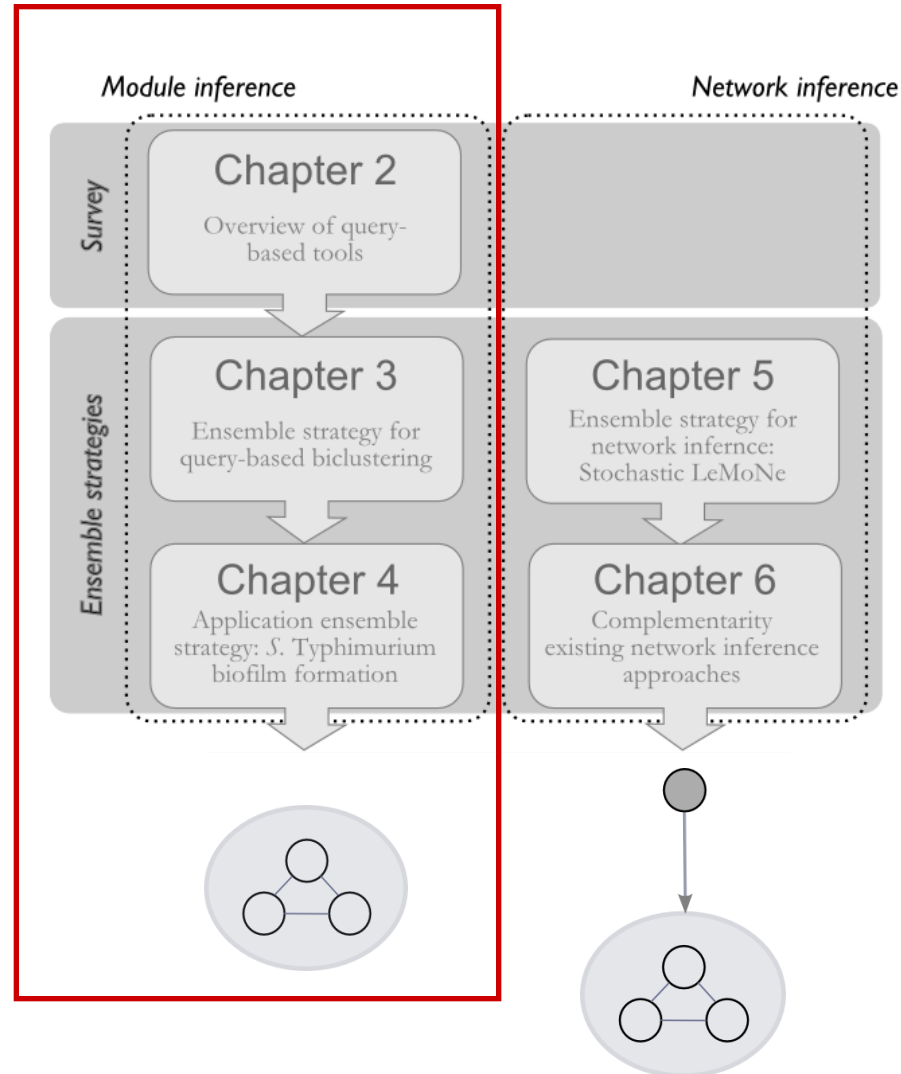
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Ensemble method for module detection

## Query-based biclustering

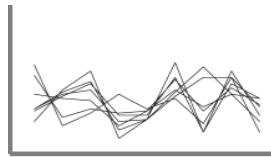
Introduction

Ensemble  
methods

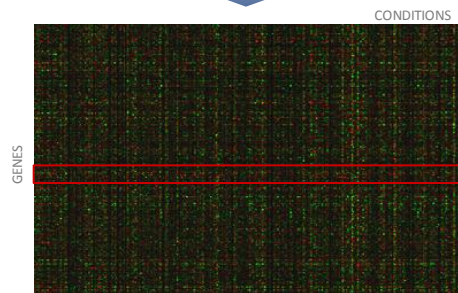
Module  
detection

Inference TRN

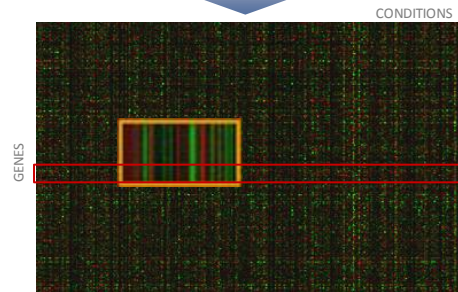
Conclusion



Genes of interest ('query/seeds')



Interrogate gene expression compendium



Find genes coexpressed with query  
and select the relevant conditions  
(‘biclustering’)

[Slide: courtesy of Lore Cloots]

# Ensemble method for module detection

## Query-based biclustering

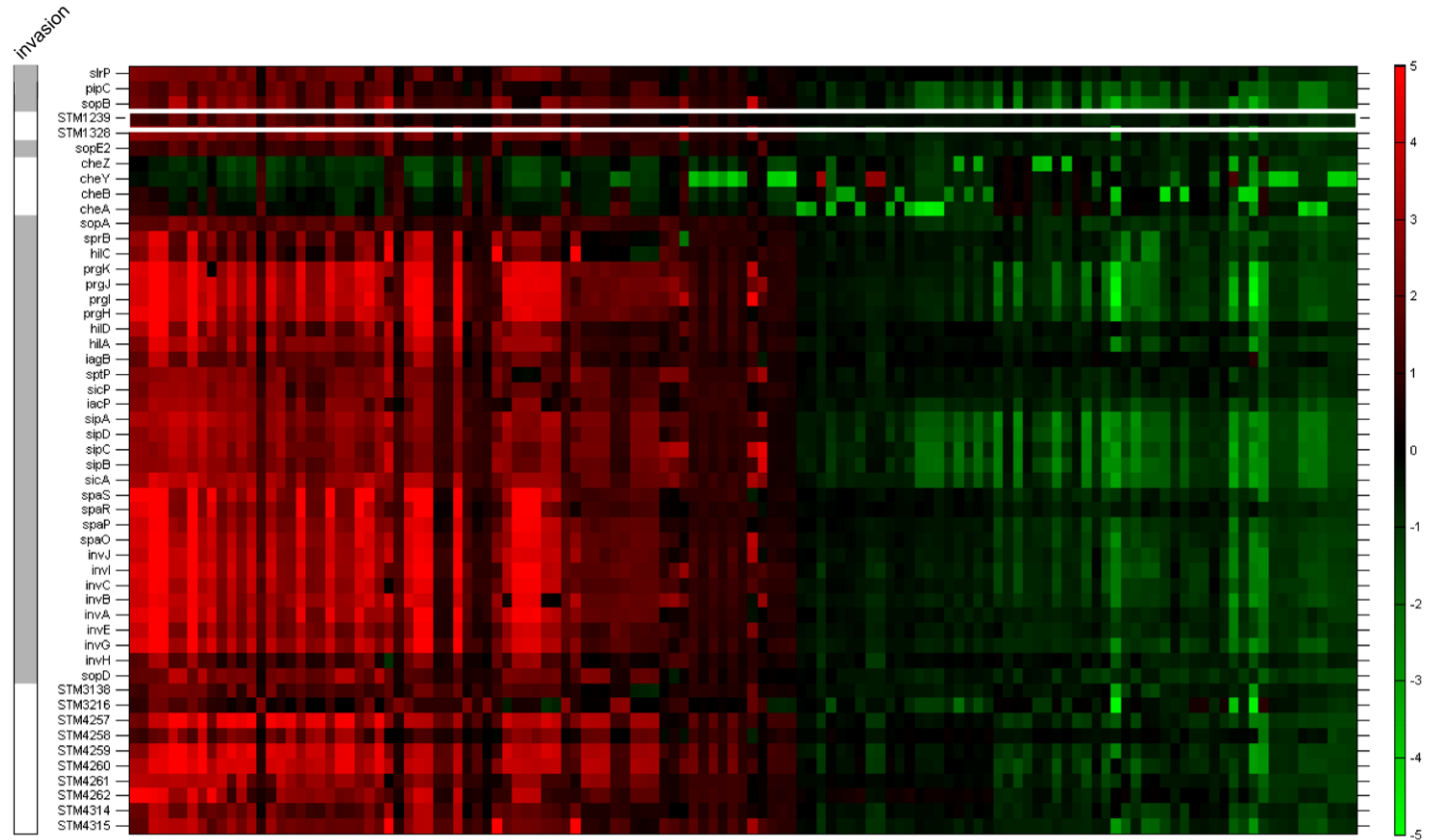
Introduction

Ensemble  
methods

Module  
detection

Inference TRN

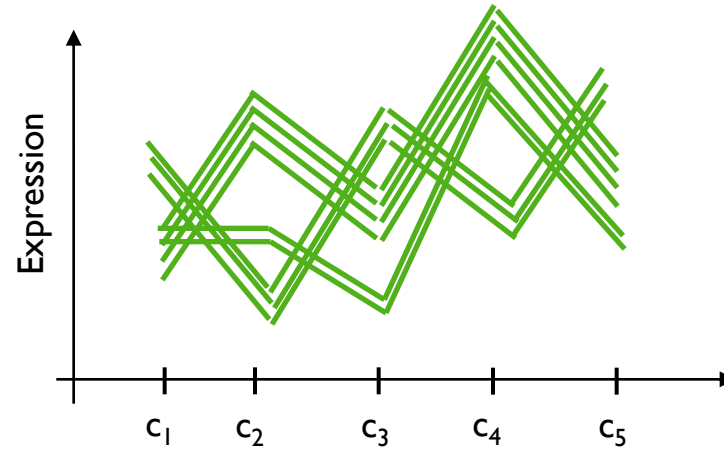
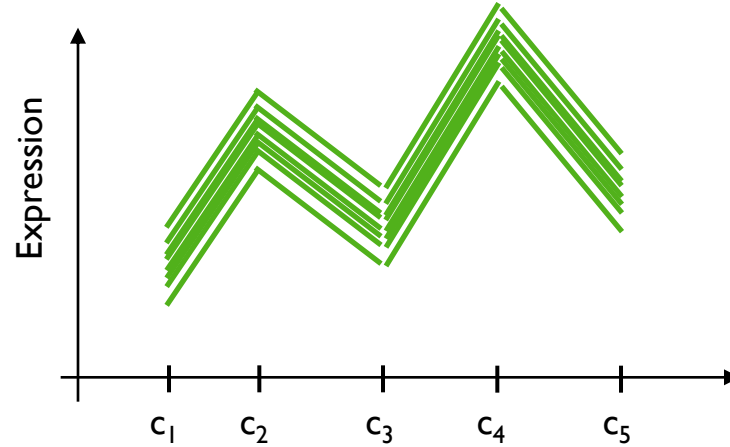
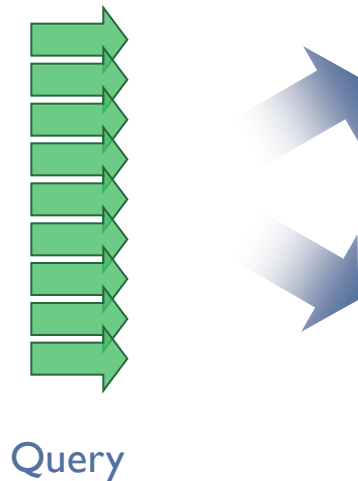
Conclusion



[Dhollander et al., 2007]  
[Thijs et al., submitted]

# Ensemble method for module detection

Problem: heterogeneity expression profiles query



Introduction

Ensemble methods

Module detection

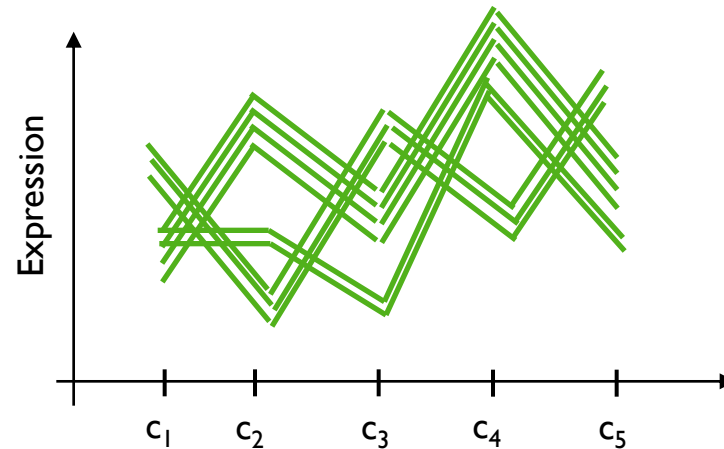
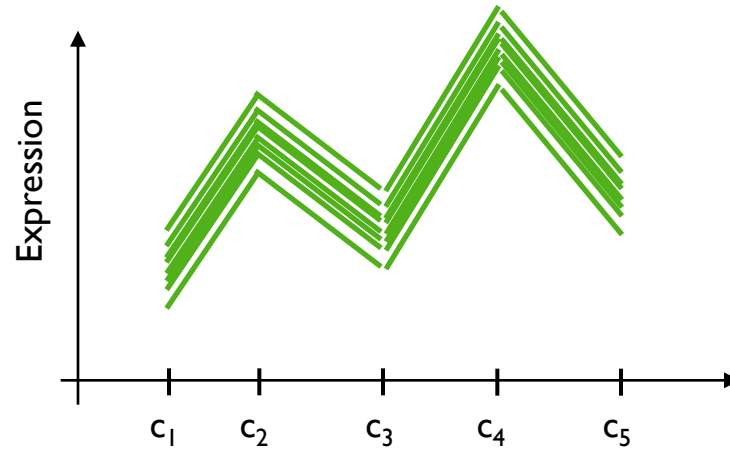
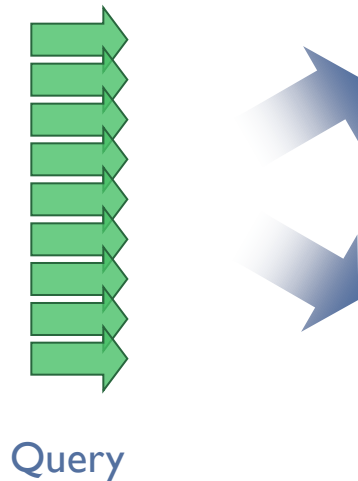
Inference TRN

Conclusion



# Ensemble method for module detection

Problem: heterogeneity expression profiles query



Introduction

Ensemble methods

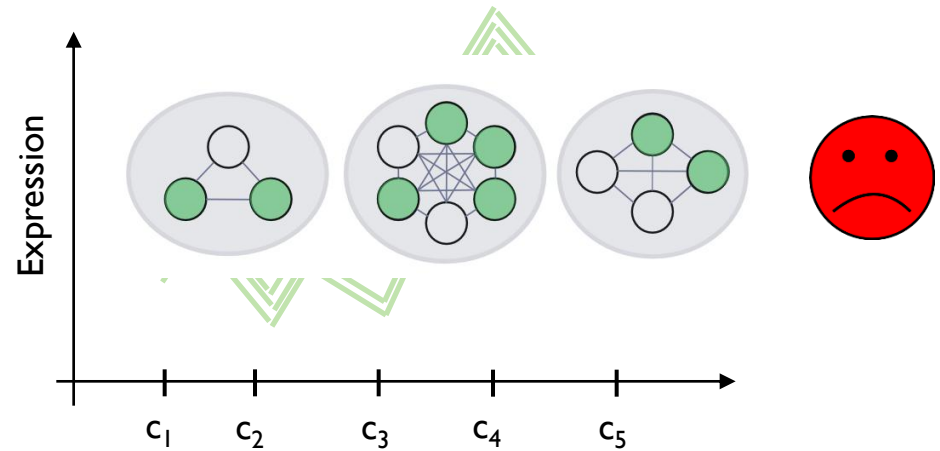
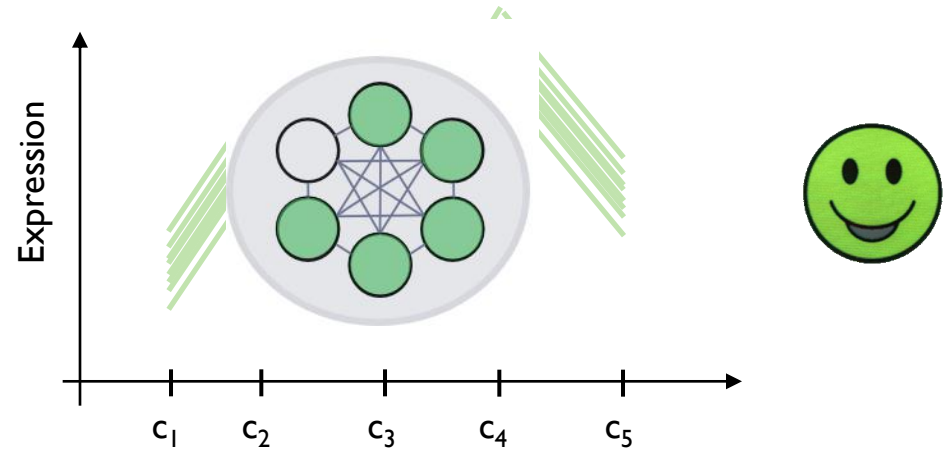
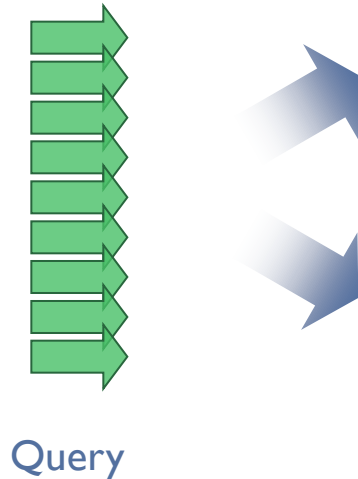
Module detection

Inference TRN

Conclusion

# Ensemble method for module detection

Problem: heterogeneity expression profiles query



e.g. experimental output

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Ensemble method for module detection

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

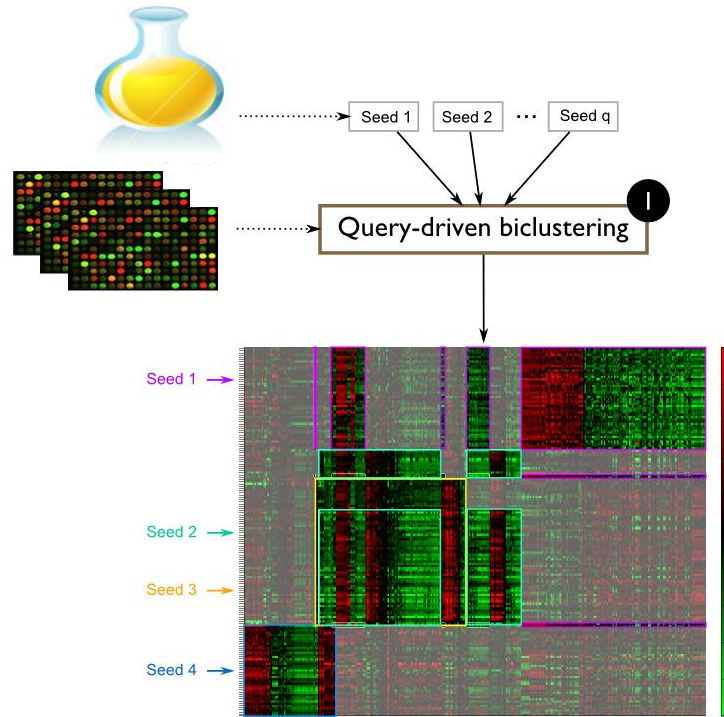
Ensemble strategy for query-driven biclustering:  
split and merge strategy



De Smet, R., Marchal, K. (2010). An ensemble method for querying gene expression compendia with experimental lists. Accepted for publication in proceedings of the IEEE International Conference on Bioinformatics and Biomedicine.

# Ensemble method for module detection

## Split and merge strategy



Introduction

Ensemble  
methods

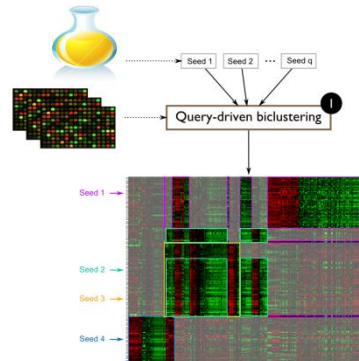
Module  
detection

Inference TRN

Conclusion

# Ensemble method for module detection

Split and merge strategy



**Ensemble**



**Consensus**

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

# Ensemble method for module detection

Biological case study:

- *E. coli* ChIP-chip experiment: **Chapter 3**
- *S. Typhimurium* biofilm assay: **Chapter 4**

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



# Ensemble method for module detection

Biological case study:

- *E. coli* ChIP-chip experiment: **Chapter 3**
- *S. Typhimurium* biofilm assay: **Chapter 4**

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

# Ensemble method for module detection

Biological case study: *E. coli* ChIP-chip experiment

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion



ChIP-chip:

- False positives
- Functional vs. non-functional binding

QDB + ensemble: **express confidence in outcome ChIP-chip**

# Ensemble method for module detection

Biological case study: *E. coli* ChIP-chip experiment

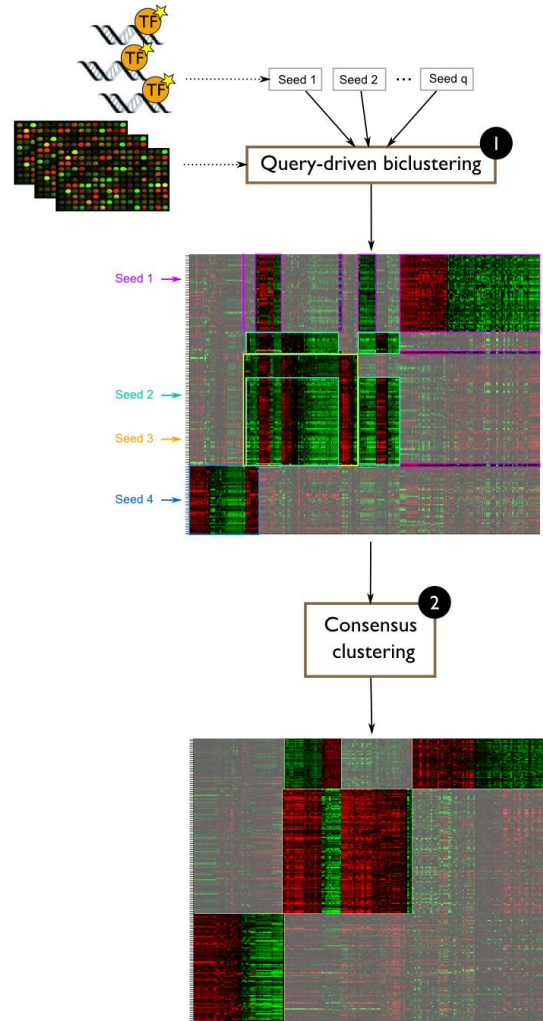
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



90 genes bound by FNR  
[Grainger et al., 2007]

61 genes

12 biclusters

# Ensemble method for module detection

Biological case study: *E. coli* ChIP-chip experiment

Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

Likely true positives/functional targets:

- Mutually coexpressed

- 1** Enrichment ChIP-chip targets

- Coexpressed with other known targets

- 2** % known FNR targets

# Ensemble method for module detection

Biological case study: *E. coli* ChIP-chip experiment

Likely true positives/functional targets:

- Mutually coexpressed

- 1** Enrichment ChIP-chip targets

- Coexpressed with other known targets

- 2** % known FNR targets

		In interesting bicluster	
	In consensus	enrich <b>1</b>	coverage <b>2</b>
Novel target	37	7	8
Known target	24	17	20
Total	61	24	28



Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Ensemble method for module detection

Biological case study: *E. coli* ChIP-chip experiment

Likely true positives/functional targets:

Pinpoints likely true positive targets of FNR

2 % known FNR targets

	In consensus	In interesting bicluster	
		enrich <b>1</b>	coverage <b>2</b>
Novel target	37	7	8
Known target	24	17	20
Total	61	24	28



# Ensemble method for module detection

## Biological case study: *E. coli* ChIP-chip experiment

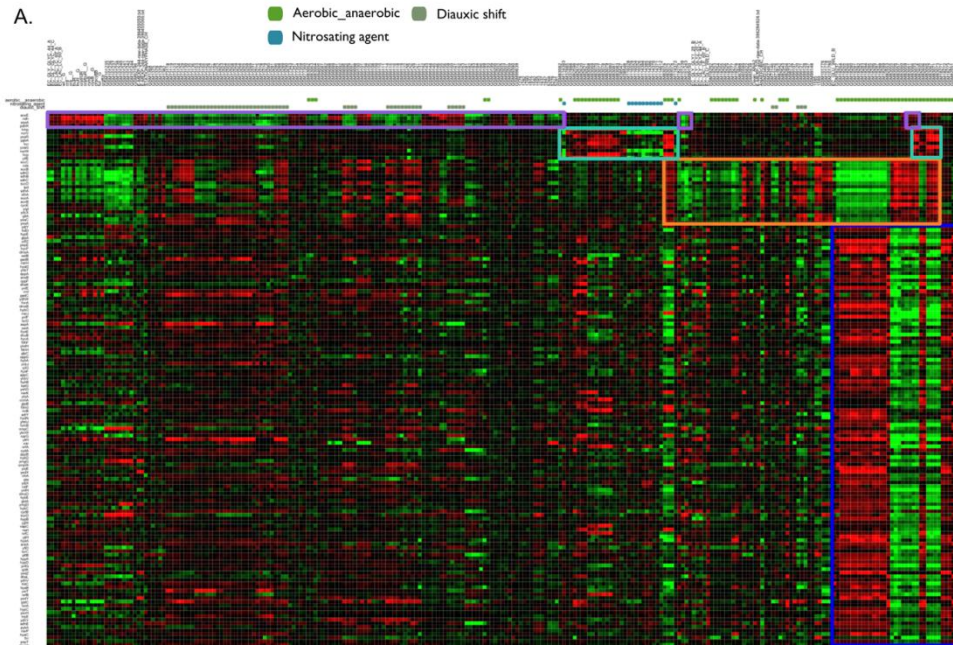
Introduction

Ensemble methods

Module detection

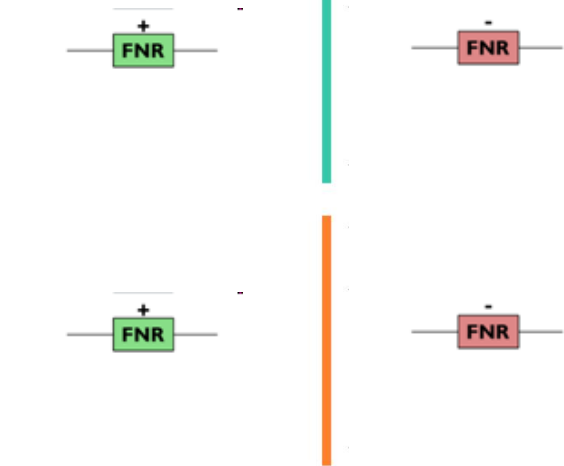
Inference TRN

Conclusion



B.

Legend:  
Bicluster 5 (purple)  
Bicluster 10 (teal)  
Bicluster 12 (blue)  
Bicluster 16 (orange)



ChIP-chip  
successful

ChIP-chip  
fails

# Ensemble method for module detection

## Biological case study: *E. coli* ChIP-chip experiment

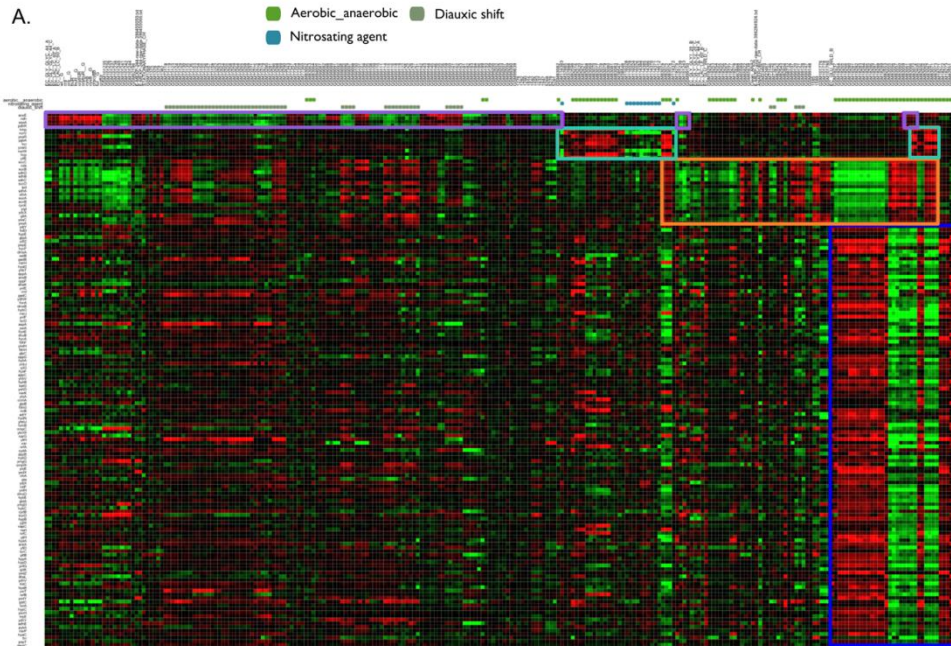
Introduction

Ensemble methods

Module detection

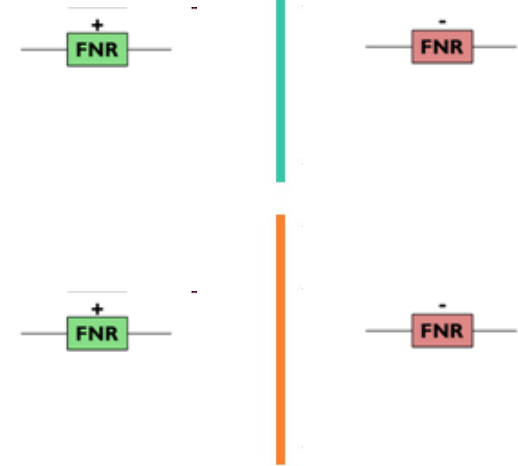
Inference TRN

Conclusion



B.

Legend:  
Bicluster 5 (purple)  
Bicluster 10 (teal)  
Bicluster 12 (blue)  
Bicluster 16 (orange)



Reveals experimental inconsistencies

High enrichment  
High coverage

ChIP-chip  
successful

Low enrichment  
High coverage

ChIP-chip  
fails

# Ensemble method for inference TRN

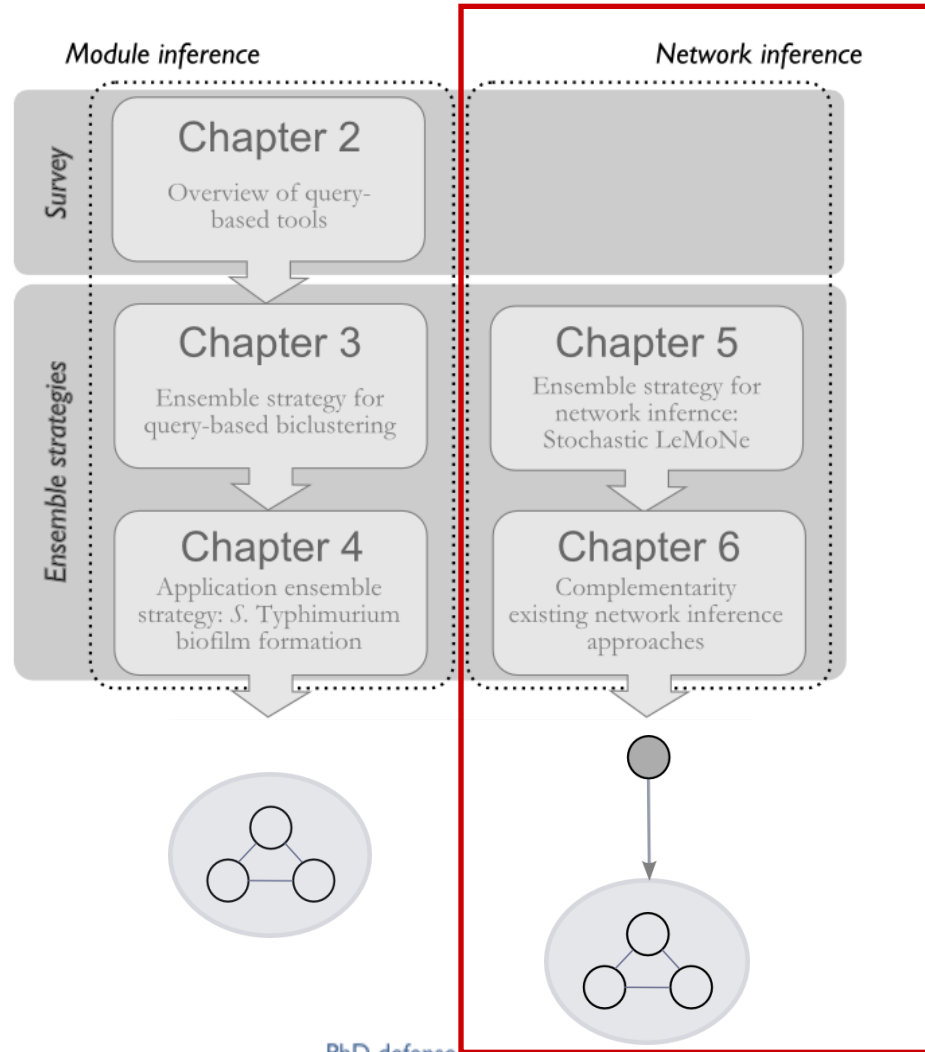
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



# Ensemble method for inference TRN

Module networks procedure [Segal et al., 2003]

Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion

**Module:** set of coexpressed genes (condition-specific)



# Ensemble method for inference TRN

Module networks procedure [Segal et al., 2003]

Introduction

Ensemble methods

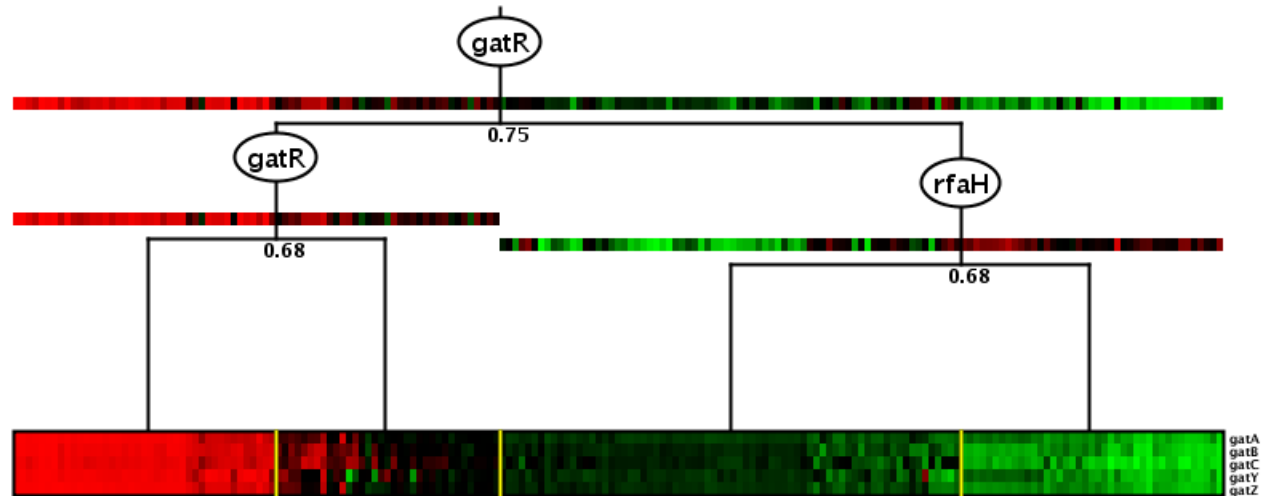
Module detection

Inference TRN

Conclusion

**Module:** set of coexpressed genes (condition-specific)

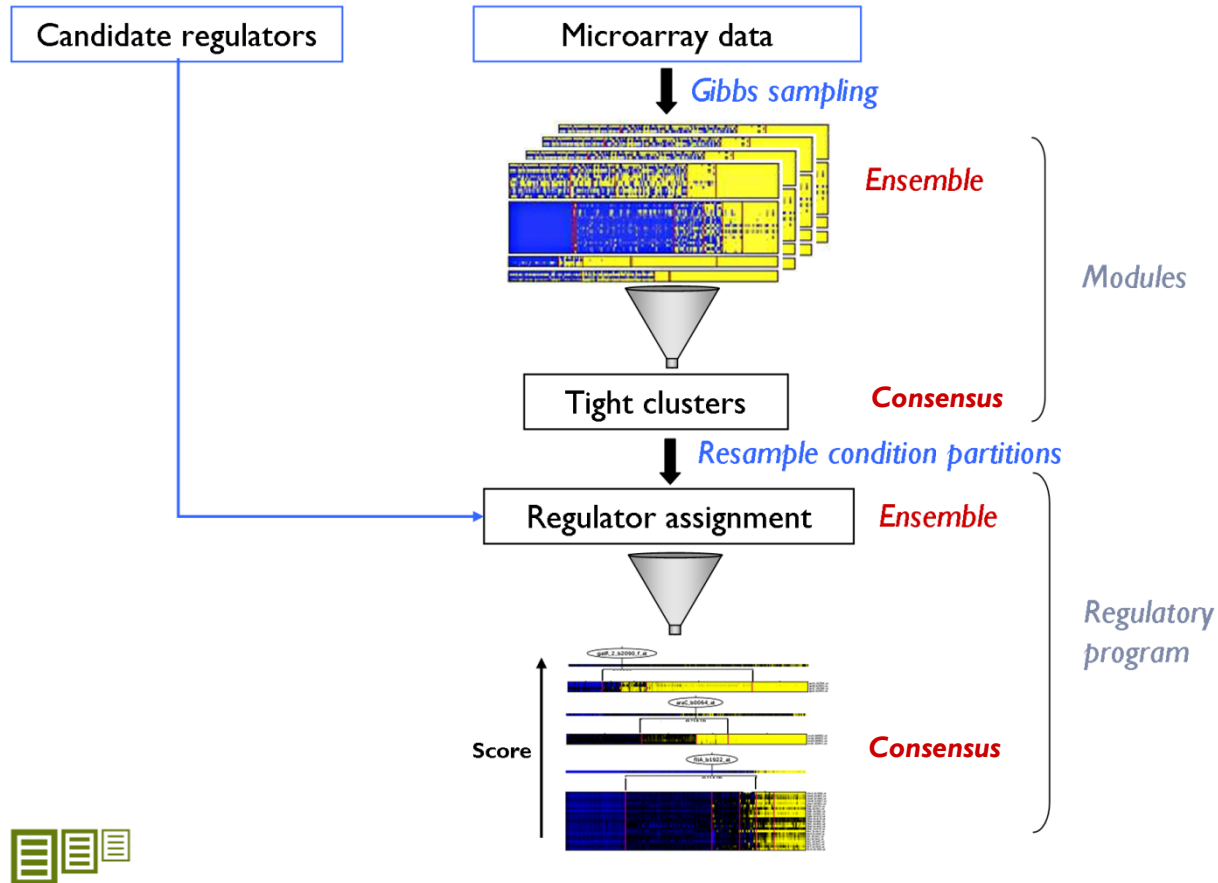
**Regulatory program:** regulators that explain expression pattern module genes





# Ensemble method for inference TRN

## Stochastic LeMoNe



Michoel, T., De Smet, R., Joshi, A., Marchal, K., Van de Peer, Y. (2009). Reverse-engineering transcriptional modules from gene expression data. *Annals of the New York Academy of Sciences*, 1158, 36-43.

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4), 490-496.



# Ensemble method for inference TRN

## Stochastic LeMoNe

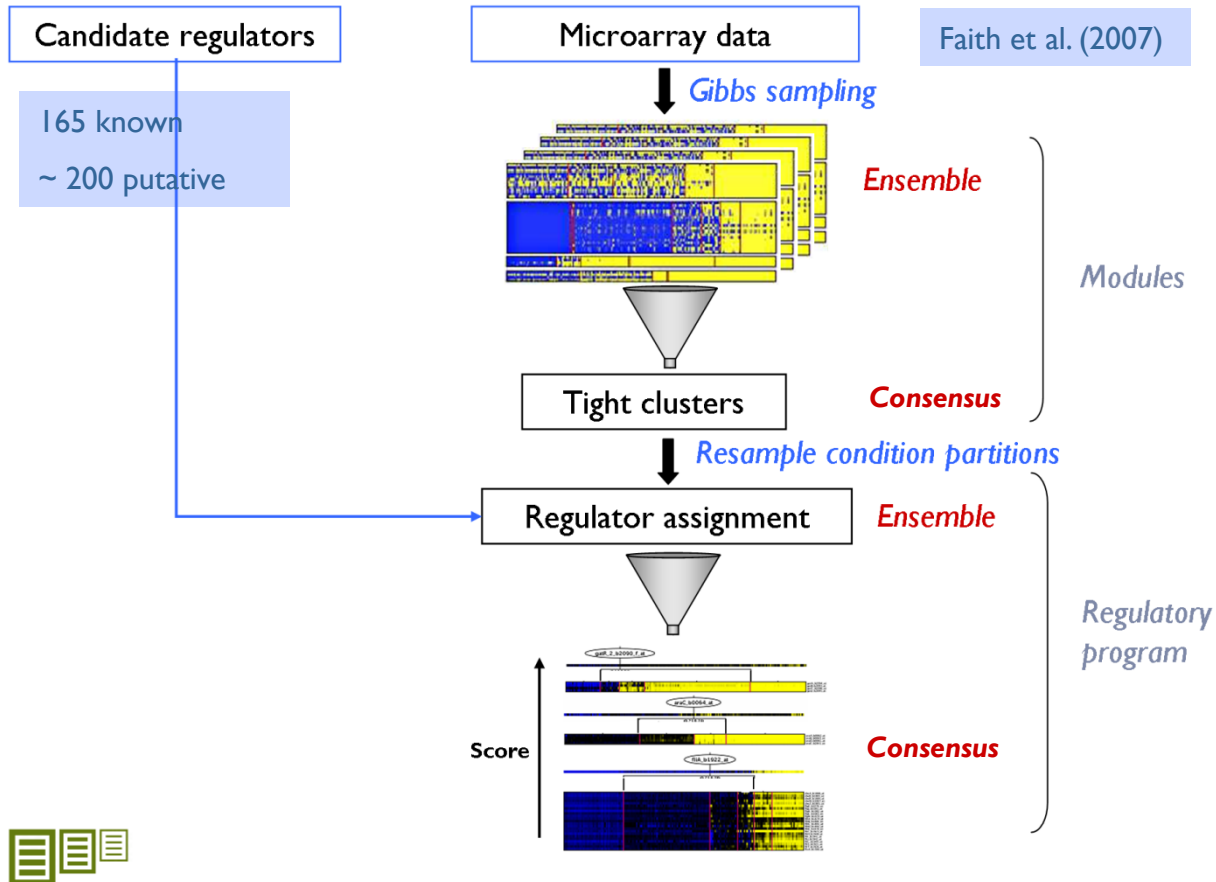
Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion



Michoel, T., De Smet, R., Joshi, A., Marchal, K., Van de Peer, Y. (2009). Reverse-engineering transcriptional modules from gene expression data. *Annals of the New York Academy of Sciences*, 1158, 36-43.

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4), 490-496.

# Ensemble method for inference TRN



Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion

- What are the consequences of using an ensemble approach?
- How does LeMoNe compare to other network inference methods?

# Ensemble method for inference TRN



Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion

- What are the consequences of using an ensemble approach?
- How does LeMoNe compare to other network inference methods?

# Ensemble method for inference TRN

## Consequences ensemble

Candidate regulators

165 known  
~ 200 putative

Microarray data

Faith et al. (2007)

Gibbs sampling



Ensemble

Modules

Tight clusters

Centroid

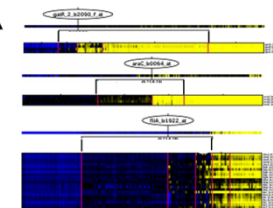
Resample condition partitions

Regulator assignment

Ensemble

Regulatory program

Score



Centroid

Ranking biologically meaningful?

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Ensemble method for inference TRN



## Consequences ensemble

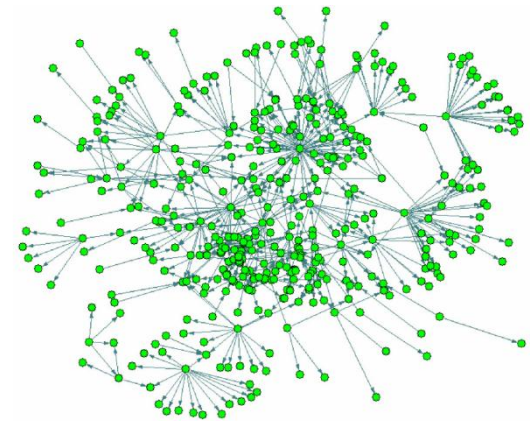
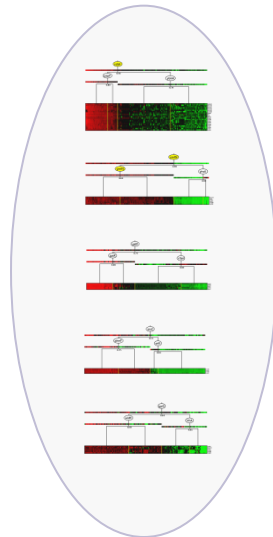
Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion



**RegulonDB**  
Escherichia coli K12 Transcriptional Network

# Ensemble method for inference TRN

## Consequences ensemble

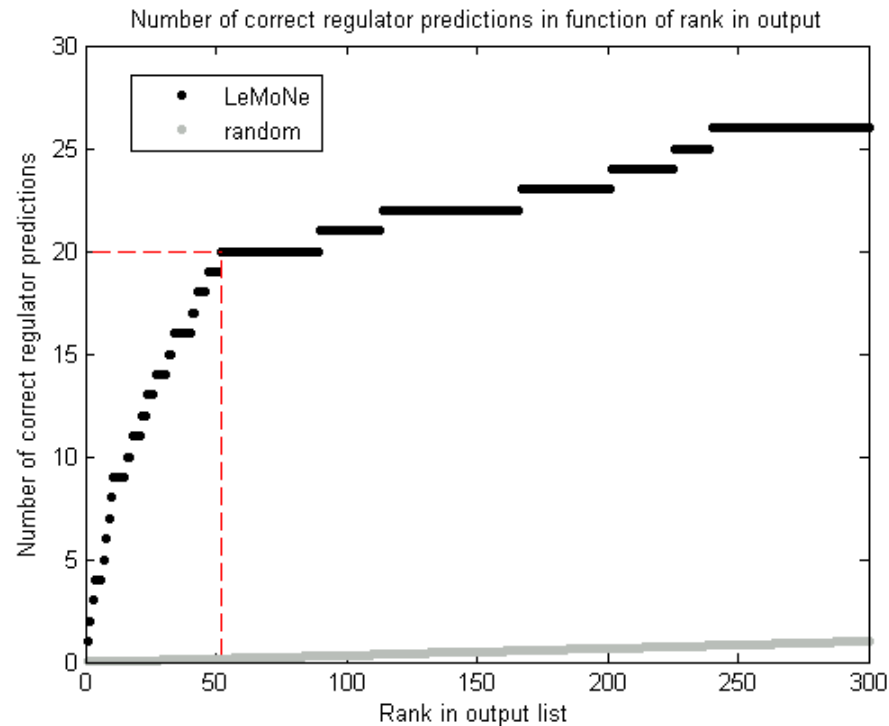
Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion



# Ensemble method for inference TRN



Introduction

Ensemble  
methods

Module  
detection

Inference  
TRN

Conclusion

- What are the consequences of using an ensemble approach?
- How does LeMoNe compare to other network inference methods?



# Ensemble method for inference TRN

Introduction

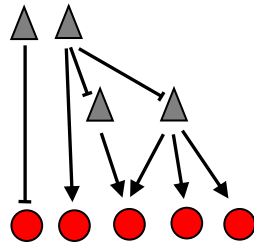
Ensemble methods

Module detection

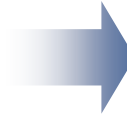
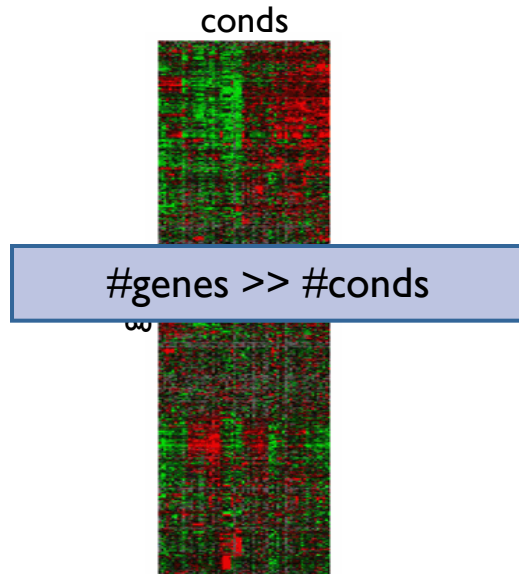
Inference TRN

Conclusion

Biological complexity



Limits data



Large search space

System heavily **underdetermined**

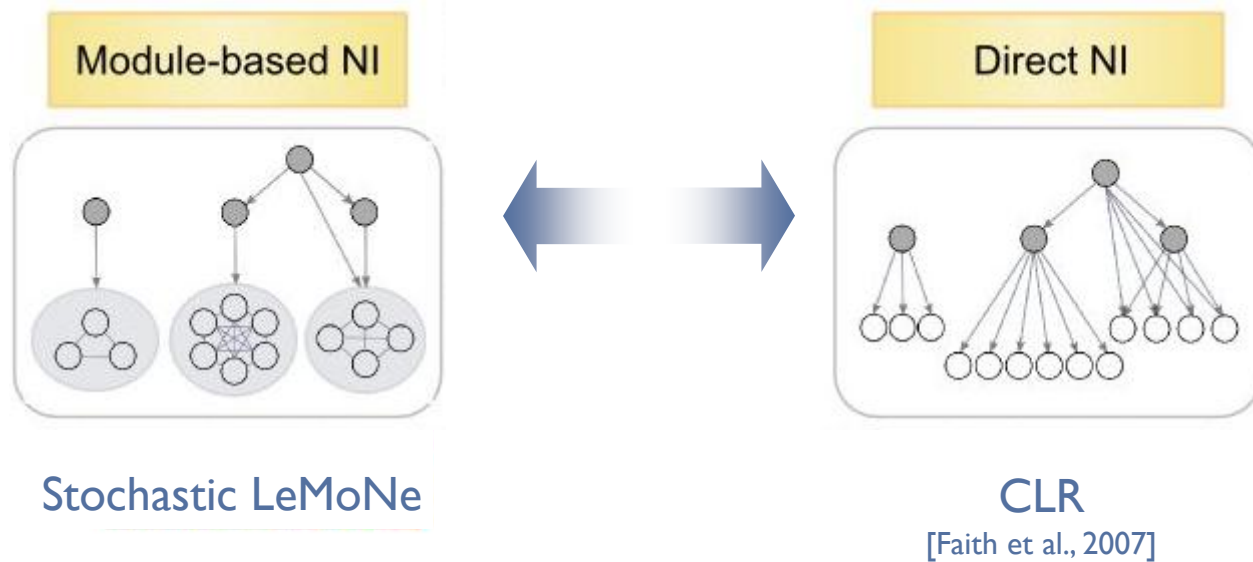


Limited number of samples

# Ensemble method for inference TRN

Reduce search space = possible number of interaction that needs to be evaluated

Simplifying biological reality



Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Ensemble method for inference TRN

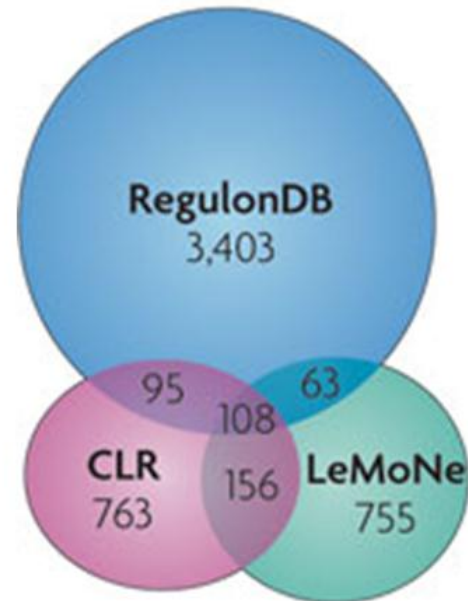
Introduction

Ensemble  
methods

Module  
detection

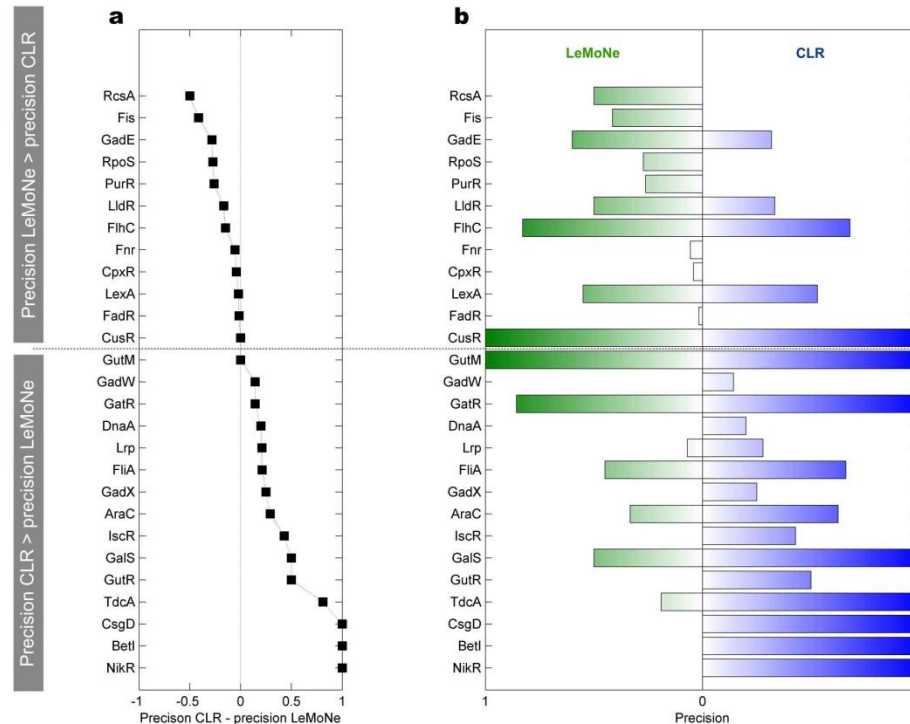
Inference  
TRN

Conclusion



# Ensemble method for inference TRN

## Comparison complementary approaches



Michoel, T., De Smet, R., Joshi, A., Van de Peer, Y., Marchal, K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Systems Biology*, 3, art.nr. 49, 49.

De Smet, R., Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8, 717-729.

Introduction

Ensemble methods

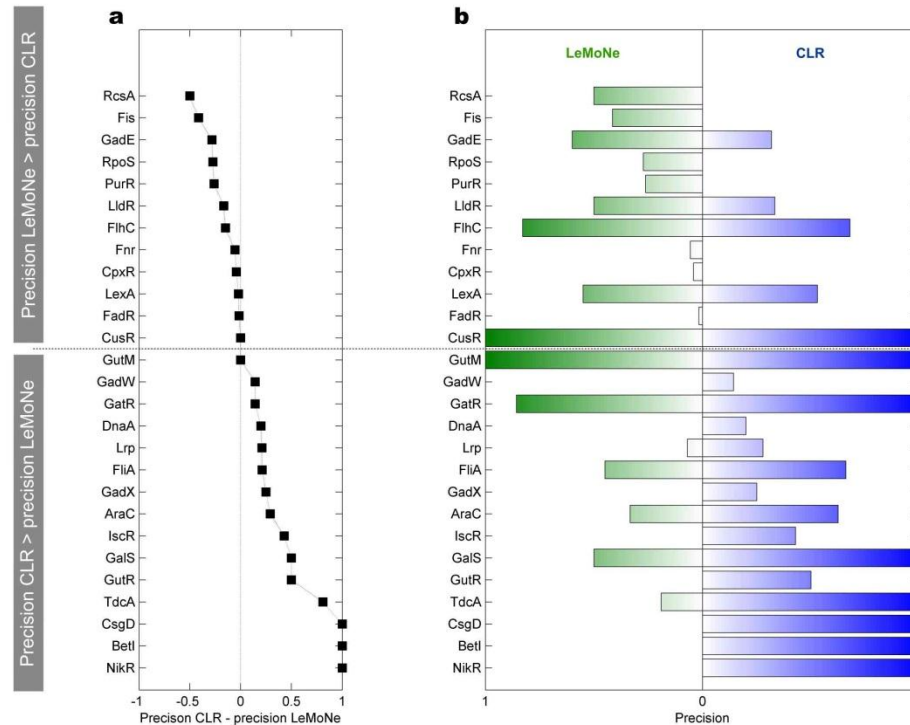
Module detection

Inference TRN

Conclusion

# Ensemble method for inference TRN

## Comparison complementary approaches



Methods show different performance at level of individual TFs

Introduction

Ensemble methods

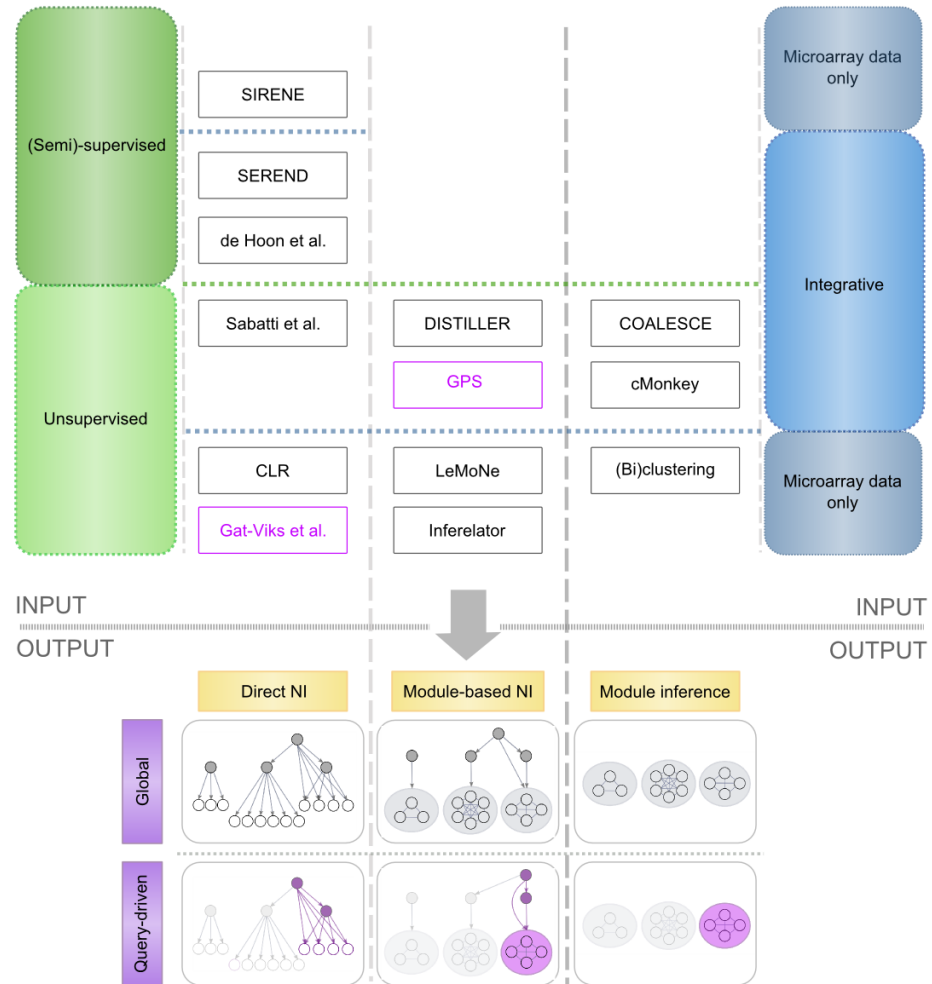
Module detection

Inference TRN

Conclusion

# Ensemble method for inference TRN

## Comparison complementary approaches



Introduction

Ensemble methods

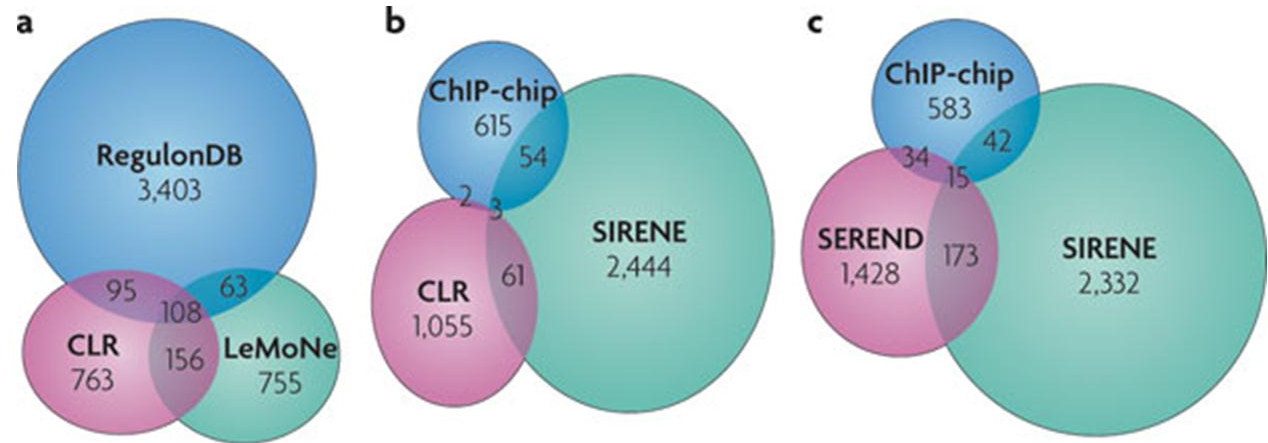
Module detection

Inference TRN

Conclusion

# Ensemble method for inference TRN

## Comparison complementary approaches



Introduction

Ensemble  
methods

Module  
detection

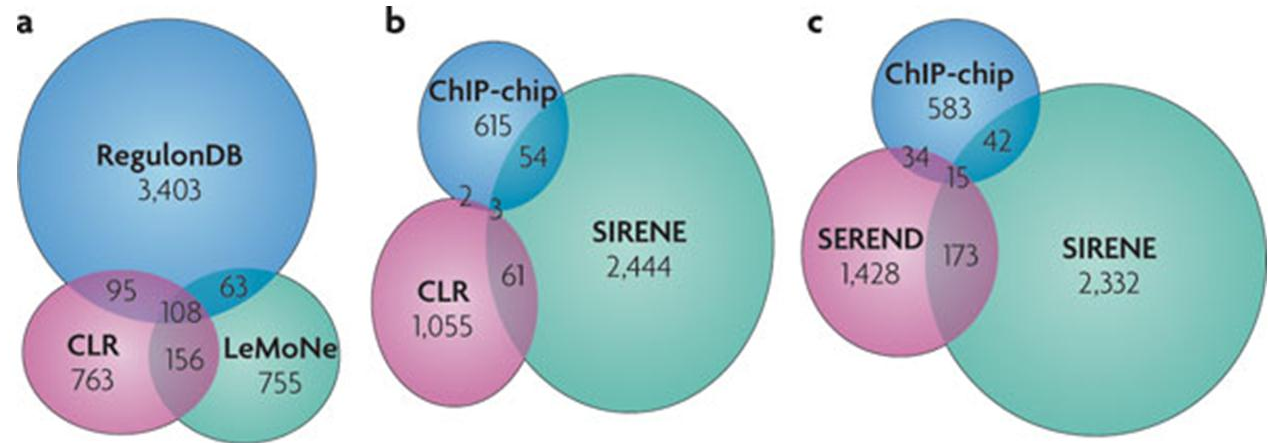
Inference  
TRN

Conclusion



# Ensemble method for inference TRN

## Comparison complementary approaches



Different strategies implemented by different NI methods results in complementary aspects being highlighted

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Conclusion

Goal PhD: improve shortcomings existing network inference methods by using ensemble strategies

- Query-based biclustering: extend application to gene lists heterogeneous in their expression profiles
- Stochastic LeMoNe: biological meaningful ranking
- Comparison NI methods: illustration complementarity → future efforts: use ensemble to exploit the complementarity of the approaches

Introduction

Ensemble methods

Module detection

Inference TRN

Conclusion

# Acknowledgements

## Promotors:

Kathleen Marchal  
Bart De Moor



Kristof Engelen  
Thomas Dhollander  
Hong Sun  
Karen Lemmens  
Inge Thijs

Jos Vanderleyden  
Kim Hermans  
Sigrid De Keersmaecker



Yves Van de Peer  
Tom Michoel  
Anagha Joshi



Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion

# Acknowledgements

How genetics works ...



Introduction

Ensemble  
methods

Module  
detection

Inference TRN

Conclusion