How to infer gene networks from gene expression data?

Jean-Philippe Vert

`Jean-Philippe.Vert@ensmp.fr`

ParisTech, Ecole des Mines de Paris
Institut Curie
INSERM U900

The Third Conference on Systems Biology, University Paris Descartes, Paris, France, April 16, 2008.

Outline

Outline

Outline

Biological networks

Data available

Biologists have collected a lot of data about proteins. e.g.,

- Gene expression measurements
- Phylogenetic profiles
- Location of proteins/enzymes in the cell

Figure 5. Phylogenetic tree congruent with cladogram in Figure 4. Redrawn from Carlson, 1995.
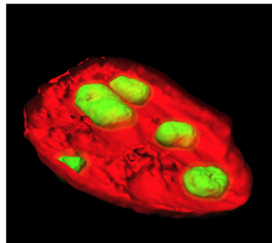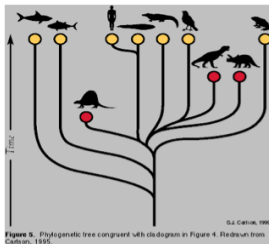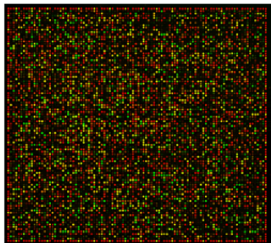
How to use this information "intelligently" to find a good function that predicts edges between nodes.

Our goal

More precisely

"De novo" inference

- Given data about individual genes and proteins
- Infer the edges between genes and proteins

"Supervised" inference

- Given data about individual genes and proteins
- and given some known interactions
- infer unknown interactions

More precisely

"De novo" inference
- Given data about individual genes and proteins
- Infer the edges between genes and proteins

"Supervised" inference
- Given data about individual genes and proteins
- and given some known interactions
- infer unknown interactions

Main messages

1. Most methods developed so far are "de novo" (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

2. Here I will focus instead on supervised methods:

3. Indeed, many real-world applications can be formulated in the supervised framework,

4. The hypothesis behind the supervised inference paradigm can be easily justified,

5. And we obtain very good results at the end.

Main messages

1. Most methods developed so far are "de novo" (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

2. Here I will focus instead on supervised methods:

3. Indeed, many real-world applications can be formulated in the supervised framework,

4. The hypothesis behind the supervised inference paradigm can be easily justified,

5. And we obtain very good results at the end.

Main messages

1. Most methods developed so far are "de novo" (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)
2. Here I will focus instead on supervised methods:
3. Indeed, many real-world applications can be formulated in the supervised framework,
4. The hypothesis behind the supervised inference paradigm can be easily justified,
5. And we obtain very good results at the end.

Main messages

1. Most methods developed so far are "de novo" (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

2. Here I will focus instead on supervised methods:

3. Indeed, many real-world applications can be formulated in the supervised framework,

4. The hypothesis behind the supervised inference paradigm can be easily justified,

5. And we obtain very good results at the end.

Main messages

1. Most methods developed so far are "de novo" (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)
2. Here I will focus instead on supervised methods:
3. Indeed, many real-world applications can be formulated in the supervised framework,
4. The hypothesis behind the supervised inference paradigm can be easily justified,
5. And we obtain very good results at the end.

De novo methods

Typical strategies

- Fit a dynamical system to time series (e.g., PDE, boolean networks, state-space models)
- Detect statistical conditional independence or dependency (Bayesian netwok, mutual information networks, co-expression)

Pros

- Excellent approach if the model is correct and enough data are available
- Interpretability of the model
- Inclusion of prior knowledge

Cons

- Specific to particular data and networks
- Needs a correct model!
- Difficult integration of heterogeneous data
- Often needs a lot of data and long computation time

De novo methods

Typical strategies

- Fit a dynamical system to time series (e.g., PDE, boolean networks, state-space models)
- Detect statistical conditional independence or dependency (Bayesian netwok, mutual information networks, co-expression)

Pros

- Excellent approach if the model is correct and enough data are available
- Interpretability of the model
- Inclusion of prior knowledge

Cons

- Specific to particular data and networks
- Needs a correct model!
- Difficult integration of heterogeneous data
- Often needs a lot of data and long computation time

Evaluation on metabolic network reconstruction

- The known metabolic network of the yeast involves 769 proteins.
- Predict edges from distances between a variety of genomic data (expression, localization, phylogenetic profiles, interactions).
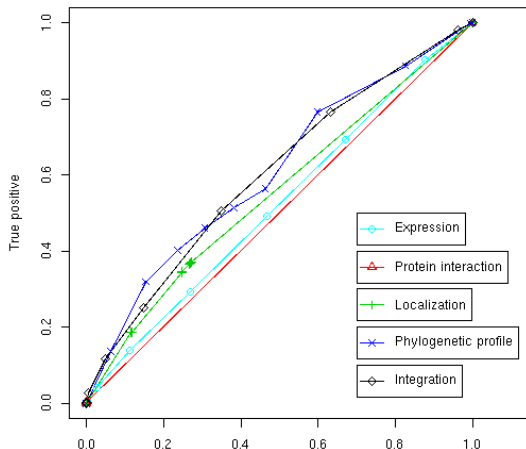
Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles

Jeremiah J. Faith[1], Boris Hayete[1], Joshua T. Thaden[2,3], Ilaria Mogno[2,4], Jamey Wierzbowski[2,5], Guillaume Cottarel[2,5], Simon Kasif[1,2], James J. Collins[1,2], Timothy S. Gardner[1,2]
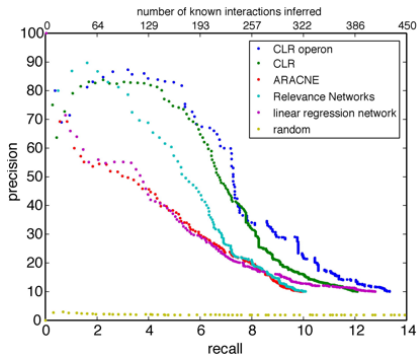
Outline

Supervised methods

Motivation

In actual applications,

- we know in advance parts of the network to be inferred
- the problem is to add/remove nodes and edges using genomic data as side information
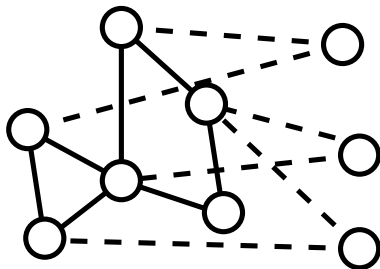
Supervised method

- Given genomic data and the currently known network...
- Infer missing edges between current nodes and additional nodes.
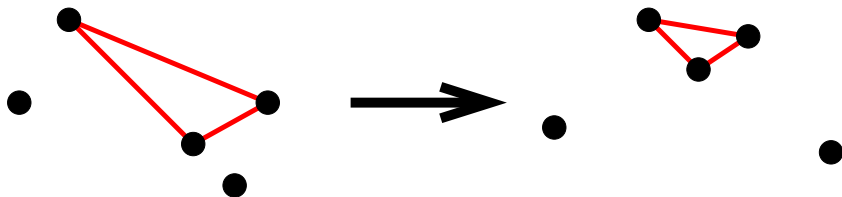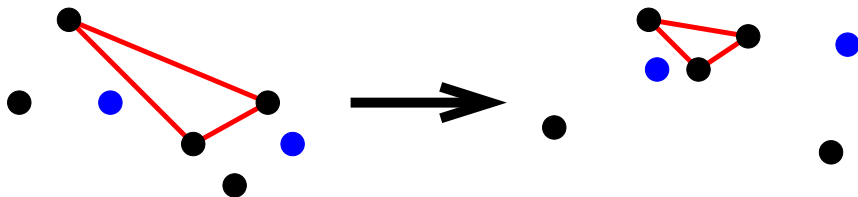
Supervised approach by Metric learning

Idea

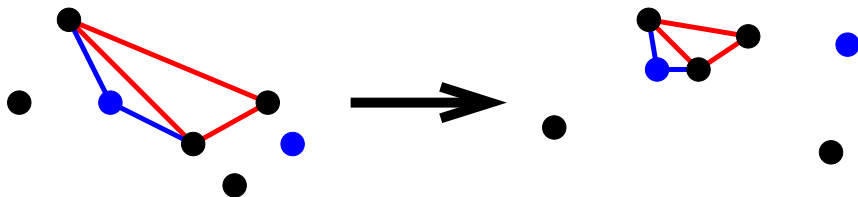- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
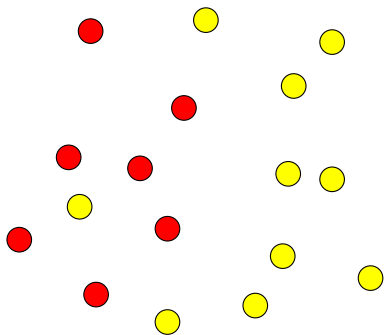- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)

Supervised approach by Metric learning

Idea

- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)
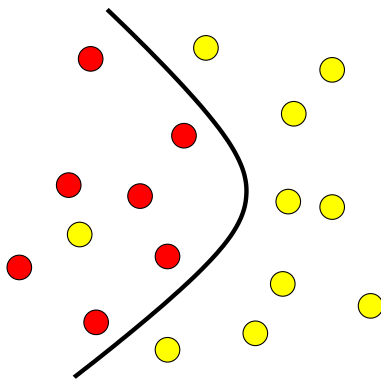
Supervised approach by Metric learning

Idea

- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
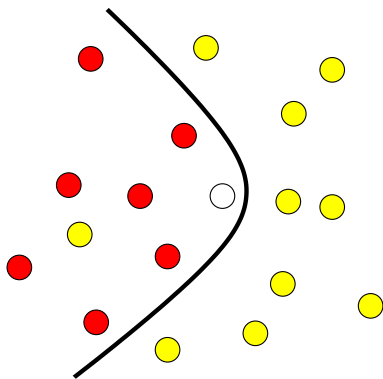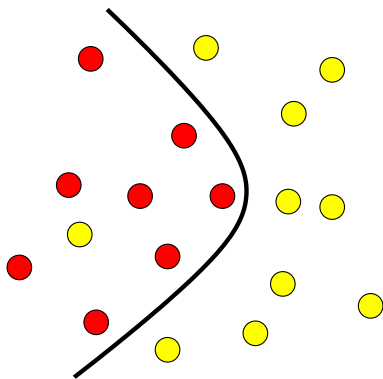- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)

Supervised approach by Metric learning

Idea

- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)

Pattern recognition

- Given a training set of patterns in two classes, learn to discriminate them
- Many algorithms (ANN, SVM, Decision tress, ...)

Pattern recognition

- Given a training set of patterns in two classes, learn to discriminate them
- Many algorithms (ANN, SVM, Decision tress, ...)

- Given a training set of patterns in two classes, learn to discriminate them
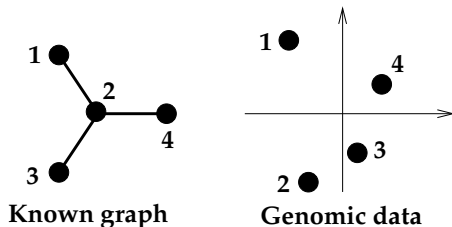- Many algorithms (ANN, SVM, Decision tress, ...)

Pattern recognition

- Given a training set of patterns in two classes, learn to discriminate them
- Many algorithms (ANN, SVM, Decision tress, ...)

Supervised inference by pattern recognition
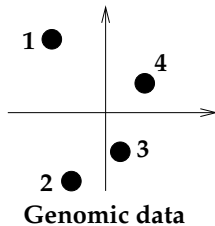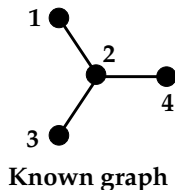
Formulation and basic issue
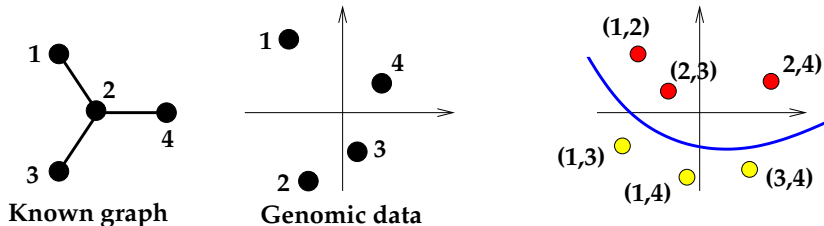
- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!

Known graph   **Genomic data**

Supervised inference by pattern recognition

Formulation and basic issue

- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!

Known graph **Genomic data**

Supervised inference by pattern recognition

Formulation and basic issue

- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!

Known graph   **Genomic data**

Tensor product SVM (Ben-Hur and Noble, 2006)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C **and** B is similar to D, **or**...
  - A is similar to D **and** B is similar to C
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}\left((a, b), (c, d)\right) = K(a, c)K(b, d) + K(a, d)K(b, c).$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized tensor product:

$$(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a).$$

Tensor product SVM (Ben-Hur and Noble, 2006)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C and B is similar to D, or...
  - A is similar to D and B is similar to C
- **Formally**, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}\left((a, b), (c, d)\right) = K(a, c)K(b, d) + K(a, d)K(b, c).$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized tensor product:

$$(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a).$$

Tensor product SVM (Ben-Hur and Noble, 2006)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C **and** B is similar to D, or...
  - A is similar to D **and** B is similar to C
- **Formally**, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}\left((a, b), (c, d)\right) = K(a, c)K(b, d) + K(a, d)K(b, c) .$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the **symmetrized tensor product**:

$$(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a) .$$

Metric learning pairwise SVM (V. et al, 2007)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - $A - B$ is similar to $C - D$, or...
  - $A - B$ is similar to $D - C$.

- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}\left((a, b), (c, d)\right) = (K(a, c) + K(b, d) - K(a, c) - K(b, d))^2 .$$

- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM

- This amounts to representing a pair (a, b) by the symmetrized difference:

$$(a, b) \rightarrow (a - b)^{\otimes 2} .$$

Metric learning pairwise SVM (V. et al, 2007)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - $A - B$ is similar to $C - D$, or...
  - $A - B$ is similar to $D - C$.
- **Formally**, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}\left((a, b), (c, d)\right) = (K(a, c) + K(b, d) - K(a, c) - K(b, d))^2 .$$

- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized difference:

$$(a, b) \rightarrow (a - b)^{\otimes 2} .$$

Metric learning pairwise SVM (V. et al, 2007)

- **Intuition**: a pair (A, B) is similar to a pair (C, D) if:
  - $A - B$ is similar to $C - D$, or...
  - $A - B$ is similar to $D - C$.
- **Formally**, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}\left((a, b), (c, d)\right) = \left(K(a, c) + K(b, d) - K(a, c) - K(b, d)\right)^2 .$$

- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized difference:

$$(a, b) \rightarrow (a - b)^{\otimes 2} .$$

Supervised inference with local models

The idea (Bleakley et al., 2007)

- Motivation: define specific models for each target node to discriminate between its neighbors and the others
- Treat each node independently from the other. Then combine predictions for ranking candidate edges.
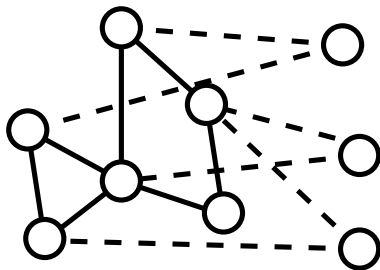
Supervised inference with local models
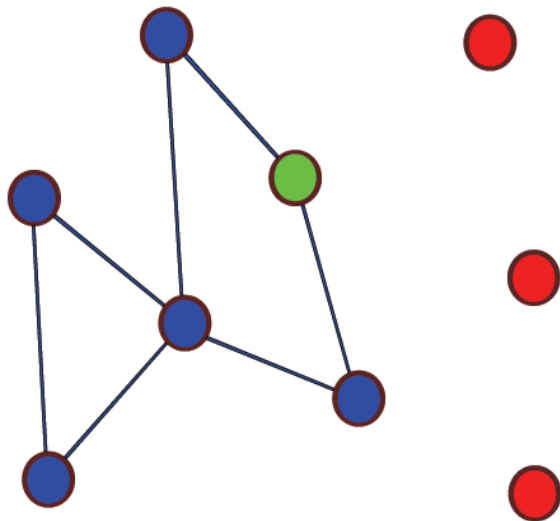
The idea (Bleakley et al., 2007)

- Motivation: define specific models for each target node to discriminate between its neighbors and the others
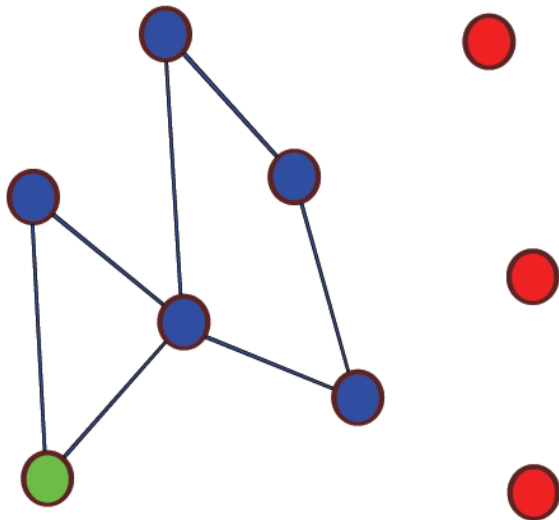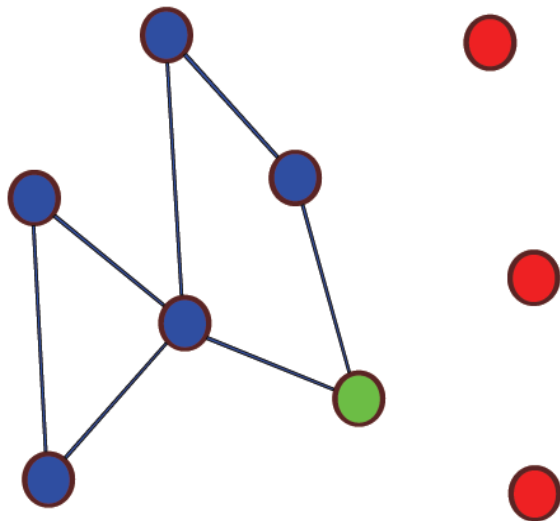- Treat each node independently from the other. Then combine predictions for ranking candidate edges.
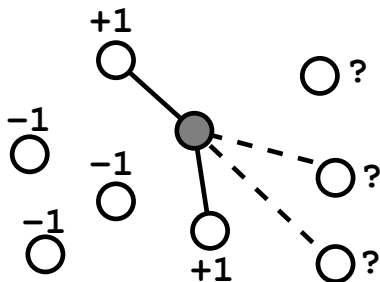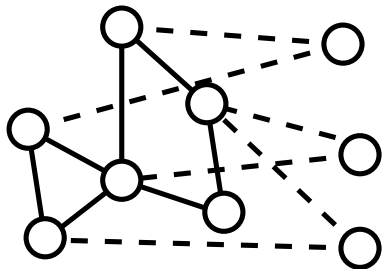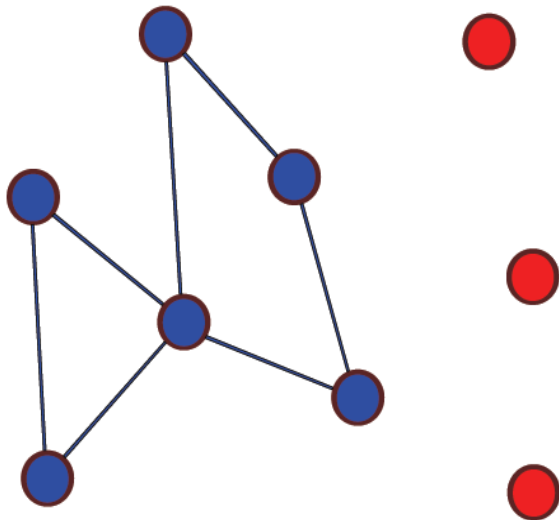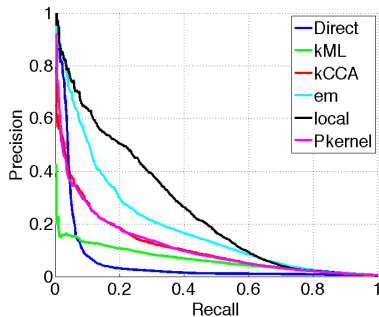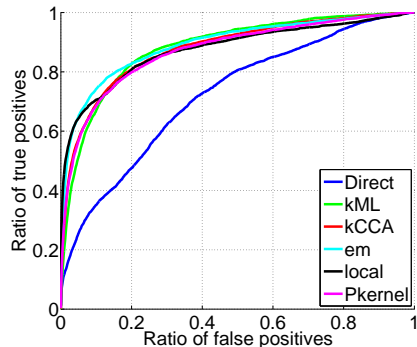
The LOCAL model

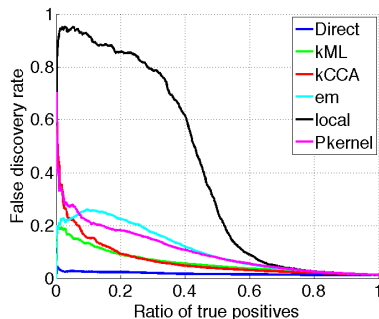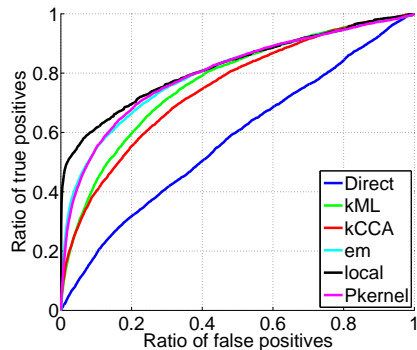# The LOCAL model
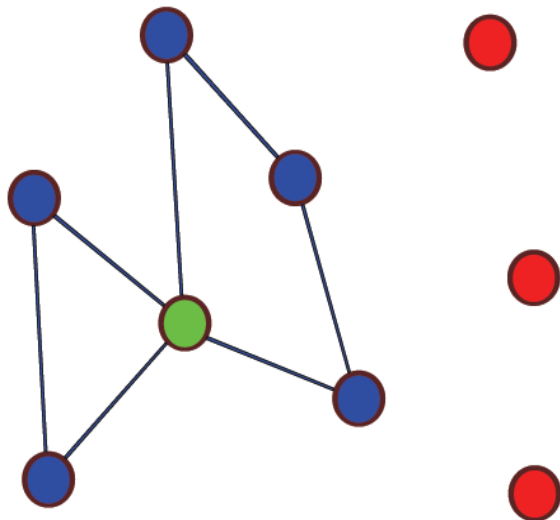
The LOCAL model

Results: protein-protein interaction (yeast)
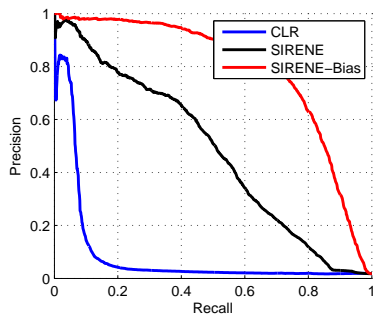

(from Bleakley et al., 2007)

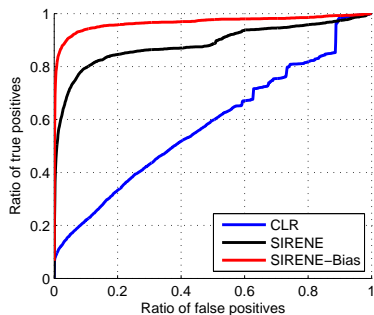# Results: metabolic gene network (yeast)

(from Bleakley et al., 2007)

Results: regulatory network (E. coli)

Method	Recall at 60%	Recall at 80%
SIRENE	**44.5%**	**17.6%**
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)

**Prediction of missing enzyme genes in a bacterial metabolic network**

Reconstruction of the lysine-degradation pathway of *Pseudomonas aeruginosa*

Yoshihiro Yamanishi[1], Hisaaki Mihara[2], Motoharu Osaki[2], Hisashi Muramatsu[3], Nobuyoshi Esaki[2], Tetsuya Sato[1], Yoshiyuki Hizukuri[1], Susumu Goto[1] and Minoru Kanehisa[1]

1 Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
2 Division of Environmental Chemistry, Institute for Chemical Research, Kyoto University, Japan
3 Department of Biology, Graduate School of Science, Osaka University, Japan

Applications: missing enzyme prediction
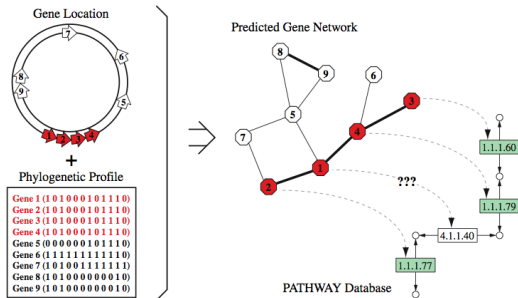
RESEARCH ARTICLE

Prediction of nitrogen metabolism-related genes in *Anabaena* by kernel-based network analysis

Shinobu Okamoto[1]*, Yoshihiro Yamanishi[1], Shigeki Ehira[2], Shuichi Kawashima[3], Koichiro Tonomura[1]** and Minoru Kanehisa[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan
[2] Department of Biochemistry and Molecular Biology, Faculty of Science, Saitama University, Saitama, Japan
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, Meguro, Japan

Determination of the role of the bacterial peptidase PepF by statistical inference and further experimental validation
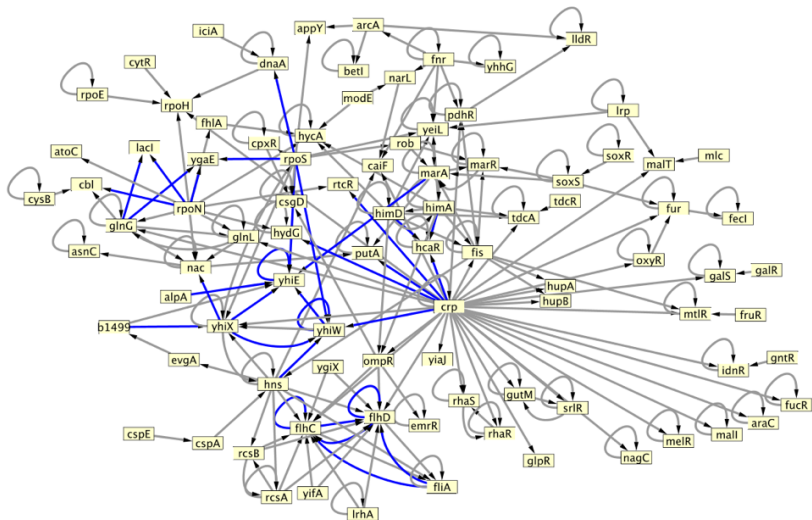
Liliana LOPEZ KLEINE[1,2], Alain TRUBUIL[1], Véronique MONNET[2]

[1]Unité de Mathématiques et Informatiques Appliquées. INRA Jouy en Josas 78352, France.
[2]Unité de Biochimie Bactérienne. INRA Jouy en Josas 78352, France.
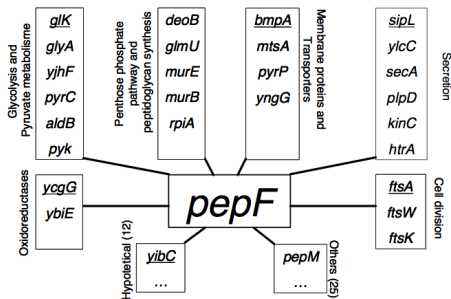
Application: predicted regulatory network (E. coli)

*Prediction at 60% precision, restricted to transcription factors (from Mordelet and V., 2008).*

Outline

Take-home messages

- When the network is known in part, supervised methods can be more adapted than unsupervised ones.
- A variety of methods have been investigated recently (metric learning, matrix completion, pattern recognition).
- The current winner on our benchmarks (metabolic, PPI and regulatory networks) is the local pattern recognition approach, which reaches high performance
- These methods:
  - work for any network
  - work with any data
  - can integrate heterogeneous data, which strongly improves performance

People I need to thank
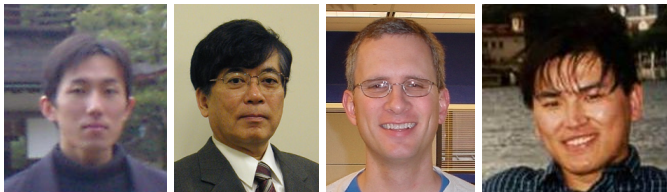
- Yoshihiro Yamanishi, Minoru Kanehisa (Univ. Kyoto): kCCA, kML
- Jian Qian, Bill Noble (Univ. Washington): pairwise SVM
- Kevin Bleakley, Gerard Biau (Univ. Montpellier), Fantine Mordelet (ParisTech/Curie): local SVM