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

Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
**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

How to use this information “intelligently” to find a good function that predicts edges between nodes.
Our goal

Jean-Philippe Vert (ParisTech-Curie)
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

<table>
<thead>
<tr>
<th>“De novo” inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given data about individual genes and proteins</td>
</tr>
<tr>
<td>Infer the edges between genes and proteins</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>“Supervised” inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given data about individual genes and proteins</td>
</tr>
<tr>
<td>and given some known interactions</td>
</tr>
<tr>
<td>infer unknown interactions</td>
</tr>
</tbody>
</table>
Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...).

Here I will focus instead on supervised methods:

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

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

And we obtain very good results at the end.
Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

Here I will focus instead on **supervised** methods:

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

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

And we obtain very good results at the end.
Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

Here I will focus instead on supervised methods:

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

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

And we obtain very good results at the end.
Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

Here I will focus instead on supervised methods:

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

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

And we obtain very good results at the end.
Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)

Here I will focus instead on supervised methods:

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

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

And we obtain very good results at the end.
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
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 network, 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 networks, 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
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

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
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)
Given a training set of patterns in two classes, learn to discriminate them

Many algorithms (ANN, SVM, Decision trees, ...)

Pattern recognition
Pattern recognition

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

Jean-Philippe Vert (ParisTech-Curie)
Pattern recognition

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

Jean-Philippe Vert (ParisTech-Curie)
Pattern recognition

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

Jean-Philippe Vert (ParisTech-Curie)
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
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!
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!
**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:

  \[
  (a, b) \rightarrow (a \otimes b) \oplus (b \otimes a) .
  \]
**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

\[
K_{TPPK} ((a, b), (c, d)) = K(a, c)K(b, d) + K(a, d)K(b, c) .
\]

- 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:

\[
(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a) .
\]
**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:

\[
(a, b) \rightarrow (a - b) \otimes^2.
\]
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:

\[
(a, b) \rightarrow (a - b)^\otimes 2.
\]
**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:

\[
(a, b) \rightarrow (a - b)^\otimes 2.
\]
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
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
The LOCAL model
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)

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall at 60%</th>
<th>Recall at 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRENE</td>
<td>44.5%</td>
<td>17.6%</td>
</tr>
<tr>
<td>CLR</td>
<td>7.5%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Relevance networks</td>
<td>4.7%</td>
<td>3.3%</td>
</tr>
<tr>
<td>ARACNe</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Bayesian network</td>
<td>1%</td>
<td>0%</td>
</tr>
</tbody>
</table>

SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)
Applications: missing enzyme prediction

Prediction of missing enzyme genes in a bacterial metabolic network

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

Yoshihiro Yamanishi¹, Hisaaki Mihara², Motoharu Osaki², Hisashi Muramatsu³, Nobuyoshi Esaki², Tetsuya Sato¹, Yoshiyuki Hizukuri¹, Susumu Goto¹ and Minoru Kanehisa¹

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

Gene Location

Phylogenetic Profile

<table>
<thead>
<tr>
<th>Gene</th>
<th>Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>101000101110</td>
</tr>
<tr>
<td>Gene 2</td>
<td>101000101110</td>
</tr>
<tr>
<td>Gene 3</td>
<td>101000101110</td>
</tr>
<tr>
<td>Gene 4</td>
<td>101000101110</td>
</tr>
<tr>
<td>Gene 5</td>
<td>00000101110</td>
</tr>
<tr>
<td>Gene 6</td>
<td>111111111110</td>
</tr>
<tr>
<td>Gene 7</td>
<td>101001111111</td>
</tr>
<tr>
<td>Gene 8</td>
<td>101000000010</td>
</tr>
<tr>
<td>Gene 9</td>
<td>101000000010</td>
</tr>
</tbody>
</table>

PATHWAY Database

1.1.1.60
1.1.1.77
4.1.1.40
1.1.1.77
1.1.1.79
Applications: missing enzyme prediction
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 KLEINE¹,², Alain TRUBUIIL¹, Véronique MONNET²

¹Unité de Mathématiques et Informatiques Appliquées. INRA Jouy en Josas 78352, France.
²Unité de Biochimie Bactérienne. INRA Jouy en Josas 78352, France.

Applications: function annotation

Jean-Philippe Vert (ParisTech-Curie)
Prediction at 60% precision, restricted to transcription factors (from Mordelet and V., 2008).
Outline

1. Inference on biological networks
2. De novo methods
3. Supervised methods
4. Conclusion
Take-home messages

- When the network is known in part, \textit{supervised} methods can be more adapted than unsupervised ones.
- A \textit{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 \textit{local pattern recognition} approach, which reaches \textit{high performance}.
- These methods:
  - work for \textit{any network}
  - work with \textit{any data}
  - can \textit{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