Inference of biological networks with supervised machine learning

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr
Mines ParisTech
Institut Curie
INSERM U900

U900 lab meeting, May 20, 2008.
Outline

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
Outline

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
Outline

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
Outline

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
Biological networks
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

Gene expression, Gene sequence, Protein localization, ...

Protein-protein interactions, Metabolic pathways, Signaling pathways, ...

Figure 4. Phylogenetic tree can be used to classify proteins (Figure A, Porphyrin from Cardio, 1989).

Jean-Philippe Vert (ParisTech-Curie)
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.
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)
Inference of biological networks
Given a training set of patterns in two classes, learn to discriminate them

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

Pattern recognition
Given a training set of patterns in two classes, learn to discriminate them

Many algorithms (ANN, SVM, Decision trees, ...
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 and graph inference

**Pattern recognition**
Associate a binary label $Y$ to each data $X$

**Graph inference**
Associate a binary label $Y$ to each pair of data $(X_1, X_2)$

**Two solutions**
- Consider each pair $(X_1, X_2)$ as a single data -> learning over pairs
- Reformulate the graph inference problem as a pattern recognition problem at the level of individual vertices -> local models
Pattern recognition

Associate a binary label $Y$ to each data $X$

Graph inference

Associate a binary label $Y$ to each pair of data $(X_1, X_2)$

Two solutions

- Consider each pair $(X_1, X_2)$ as a single data -> learning over pairs
- Reformulate the graph inference problem as a pattern recognition problem at the level of individual vertices -> local models
Pattern recognition for pairs

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!
Pattern recognition for pairs

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!
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!
Pattern recognition for pairs

Representing a pair as a vector

- Each individual protein is represented by a vector $v \in \mathbb{R}^p$
- We must represent a pair of proteins $(u, v)$ by a vector $\psi(u, v) \in \mathbb{R}^q$ in order to estimate a linear classifier
- Question: how build $\psi(u, v)$ from $u$ and $v$?
A simple idea is to concatenate the vectors $u$ and $v$ to obtain a $2p$-dimensional vector of $(u, v)$:

$$
\psi(u, v) = u \oplus v = \begin{pmatrix} u \\ v \end{pmatrix}.
$$

Problem: a linear function then becomes additive...

$$
f(u, v) = w^T \psi(u, v) = w_1^T u + w_2^T v.
$$
Representing a pair

Direct sum

- A simple idea is to **concatenate** the vectors $u$ and $v$ to obtain a $2p$-dimensional vector of $(u, v)$:

$$\psi(u, v) = u \oplus v = \left( \begin{array}{c} u \\ v \end{array} \right).$$

- **Problem**: a linear function then becomes **additive**...

$$f(u, v) = w^T \psi(u, v) = w_1^T u + w_2^T v.$$
Representing a pair

**Direct product**

- Alternatively, make the **direct product**, i.e., the $p^2$-dimensional vector whose entries are all products of entries of $u$ by entries of $v$:

  \[ \psi(u, v) = u \otimes v \]

- **Problem**: can get really large-dimensional...

- **Good news**: inner product factorizes:

  \[ (u_1 \otimes v_1)^\top (u_2 \otimes v_2) = (u_1^\top u_2) \times (v_1^\top v_2) , \]

  which is good for algorithms that use only inner products (SVM...)

Jean-Philippe Vert (ParisTech-Curie)
Representing a pair

Direct product

- Alternatively, make the **direct product**, i.e., the $p^2$-dimensional vector whose entries are all products of entries of $u$ by entries of $v$:

  $$\psi(u, v) = u \otimes v$$

- **Problem**: can get really large-dimensional...

- **Good news**: inner product factorizes:

  $$(u_1 \otimes v_1)\top (u_2 \otimes v_2) = (u_1\top u_2) \times (v_1\top v_2),$$

  which is good for algorithms that use only inner products (SVM...)

Jean-Philippe Vert (ParisTech-Curie)
Representing a pair

**Direct product**

- Alternatively, make the direct product, i.e., the $p^2$-dimensional vector whose entries are all products of entries of $u$ by entries of $v$:
  \[ \psi(u, v) = u \otimes v \]

- **Problem**: can get really large-dimensional...
- **Good news**: inner product factorizes:
  \[ (u_1 \otimes v_1) \top (u_2 \otimes v_2) = (u_1 \top u_2) \times (v_1 \top v_2), \]

  which is good for algorithms that use only inner products (SVM...)

Jean-Philippe Vert (ParisTech-Curie)
Other representations for pair

**Symmetric tensor product** (Ben-Hur and Noble, 2006)

\[ \psi(u, v) = (u \otimes v) + (v \otimes u). \]

**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 \(C\)

**Metric learning** (V. et al, 2007)

\[ \psi(u, v) = (u - v)^{\otimes 2}. \]

**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 \(D - C\).
Other representations for pair

**Symmetric tensor product (Ben-Hur and Noble, 2006)**

\[ \psi(u, v) = (u \otimes v) + (v \otimes u) . \]

**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 \(C\)

**Metric learning (V. et al, 2007)**

\[ \psi(u, v) = (u - v)^{\otimes 2} . \]

**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 \(D - C\).
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
Outline

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
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)*
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
Applications: missing enzyme prediction
Applications: missing enzyme prediction

RESEARCH ARTICLE

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

Shinobu Okamoto\textsuperscript{1,*}, Yoshihiro Yamanishi\textsuperscript{1}, Shigeki Ehira\textsuperscript{2}, Shuichi Kawashima\textsuperscript{3}, Koichiro Tonomura\textsuperscript{1,**} and Minoru Kanehisa\textsuperscript{1}

\textsuperscript{1} Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan
\textsuperscript{2} Department of Biochemistry and Molecular Biology, Faculty of Science, Saitama University, Saitama, Japan
\textsuperscript{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\textsuperscript{1,2}, Alain TRUBUI\textsuperscript{1}, Véronique MONNET\textsuperscript{2}

\textsuperscript{1}Unité de Mathématiques et Informatiques Appliquées. INRA Jouy en Josas 78352, France.
\textsuperscript{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

1. Inference of biological networks
2. Supervised methods
3. Applications
4. Conclusion
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