On Learning Gene Regulatory Networks with Only Positive Examples

Luigi Cerulo, University of Sannio, Italy
Michele Ceccarelli, University of Sannio, Italy
Outline

• Supervised inference of gene regulatory networks
• The positive only problem
• Negative selection approaches
• Effect on prediction accuracy
• Conclusions and future directions
Gene Regulatory Network (GRN)

The network of transcription dependences among genes of an organism, known as transcription factors, and their binding sites.
Gene Regulatory Network (GRN)

- A gene regulatory network can be represented as a graph \( G = (\text{Vertices}, \text{Edges}) \)
- Vertices = Genes
- Edges = Interactions
Inference of Gene regulatory networks

\[ G_i = \{e_1, e_2, e_3, \ldots, e_n\} \]
GRN
unsupervised inference

• Correlation models (eg. Mutual information)
• Bayesian Network
• Boolean networks
• ODE
• ...

Friday, October 15, 2010
GRN supervised Inference

- Part of the network is known in advance from public databases (Eg. RegulonDB)
GRN supervised Inference

\[ G_i = \{e_1, e_2, e_3, \ldots, e_n\} \]

\[ T = \{(G_1, G_2), (G_2, G_3), (G_6, G_7), (G_7, G_8)\} \]

Binary classifier (SVM, Decision Tree, Neural Networks,...)
Related work

SIRENE: supervised inference of regulatory networks
Fantine Mordelet$^{1,2,3,4,*}$ and Jean-Philippe Vert$^{1,2,3}$

$^1$Ecole des Mines de Paris, ParisTech, 35 rue Saint-Honoré, Fontainebleau F-77300, $^2$Institut Curie, Paris F-75248, $^3$INSERM, U900, Paris F-75248 and $^4$CREST, INSEE, 3 av. Pierre Larousse, Malakoff, F-92240 France
• SIRENE approach

• trains an SVM classifier for each gene and predicts which genes are regulated by that gene

• combines all predicted regulations to obtain the full regulatory network
Compared with unsupervised methods (Mordelet and Vert, 2008)
Supervised learning
Supervised learning
Supervised learning
Supervised learning
Supervised learning

\[ f(x) \]
Supervised learning with unlabeled data
Supervised learning with unlabeled data
Supervised learning with unlabeled data
Supervised learning with unlabeled data

PU-learning
Supervised learning of gene regulatory networks

Is this a negative example?

Is this a negative example?
Training set

Labeled

P

Q

Unlabeled

N

Positive

Negative

% of Known Positives

\[
\frac{|P|}{|P \cup Q|}
\]
Effect of PU-learning
E.coli dataset [J.J. Faith et al., 2007]
Reliable negative selection

PU-learning
Reliable negative selection
Reliable negative selection
Reliable negative selection in text mining


• Yu et al. PEBL: Positive Example Based Learning for Web Page Classification Using SVM, in KDD 2002

• Denis et al. Text classification from positive and unlabeled Examples, in IPMU 2002
Methods based on reliable negative selection

- Original training set
  - Labeled: P
  - Unlabeled: Q

- Negative selection heuristic

- New training set
  - Labeled: P
  - Unlabeled: RN
Quality of RN

- RN could be contaminated with positives embedded in unlabeled data
- The fraction of positive contamination is the ratio between the number of positives in RN and the total number of unknown positives $|Q|$
Effect of positive contamination

E.coli dataset [J.J. Faith et al., 2007]
Effect of positive contamination

E.coli dataset [J.J. Faith et al., 2007]
Network topology based heuristics
Network motifs

Network motifs are small connected subnetworks a network exhibits in a significant higher or lower occurrences than would be expected just by chance.

Network Motifs
Heuristic

- For each three genes sub networks $T$:
- If matches a network motifs $M$ then considers all connections not present in $M$ as negatives
• For each three genes sub networks $T$:
• If matches a network motifs $M$ then considers all connections not present in $M$ as negatives
Network Motifs Heuristic

- For each three genes sub networks $T$:  
- If matches a network motifs $M$ then considers all connections not present in $M$ as negatives
Network Motifs Heuristic

- For each three genes sub networks T:
- If matches a network motifs M then considers all connections not present in M as negatives
MOTIF selection performance

E.coli dataset [J.J. Faith et al., 2007 and RegulonDB]
MOTIF selection performance

E.coli dataset [J.J. Faith et al., 2007 and RegulonDB]
Effect of positive contamination

E.coli dataset [J.J. Faith et al., 2007]

% of known positives

F-Measure

positive contamination = 0

positive contamination = 1
( PU-learning )
Effect of positive contamination

E.coli dataset [J.J. Faith et al., 2007]

% of known positives
Scale free networks

Albert-László Barabási and Zoltán N. Oltvai
Network biology: Understanding the cell’s functional organization
Hierarchical networks

Hong-Wu Ma, Jan Buer, and An-Ping Zeng
Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach
BMC Bioinformatics 2004 5:199
Experimental data

- 445 Affymetrix Antisense2 microarray expression profiles for 4345 genes of E.coli [J.J. Faith et al., 2007]
- Data were standardized (i.e. zero mean unit standard deviation)
- Regulations extracted from RegulonDB (v. 5) between 154 Transcription Factors and 1211 genes
Summary and conclusions

- Learning gene regulations is affected by the problem of learning from positive only data
- At least for E.coli
  - The study of positive contamination shows that there is room for new heuristics
  - Topology based heuristics (eg. motifs) have shown promising results.
- Open issues arise on higher level organisms where gene interactions are more complex