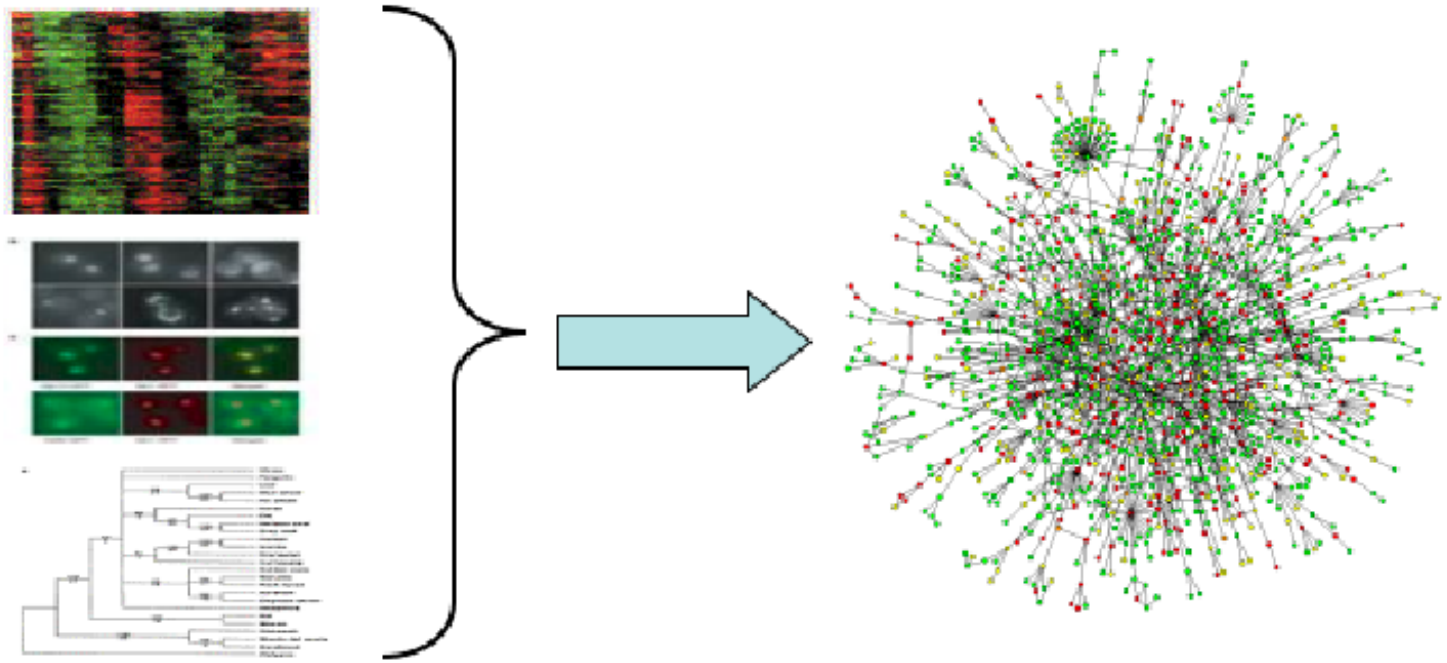
Supervised inference of biological networks from heterogeneous genomic data

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Centre for Computational Biology
Ecole des Mines de Paris, ParisTech

ENFIN NoE meeting, Evry, France, February 2nd, 2007

Motivation



Data

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

Graph

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

Strategies

Unsupervised approaches

The graph is completely unknown

- model-based approaches : Bayes nets, dynamical systems,..
- similarity-based : connect similar nodes

Supervised approaches

Part of the graph is known in advance

- Prior knowledge in model-based approaches
- Statistical / Machine learning approaches: learn from the known subnetwork a rule that can predict edges from genomic data

Strategies

Unsupervised approaches

The graph is completely unknown

- model-based approaches : Bayes nets, dynamical systems,..
- similarity-based : connect similar nodes

Supervised approaches

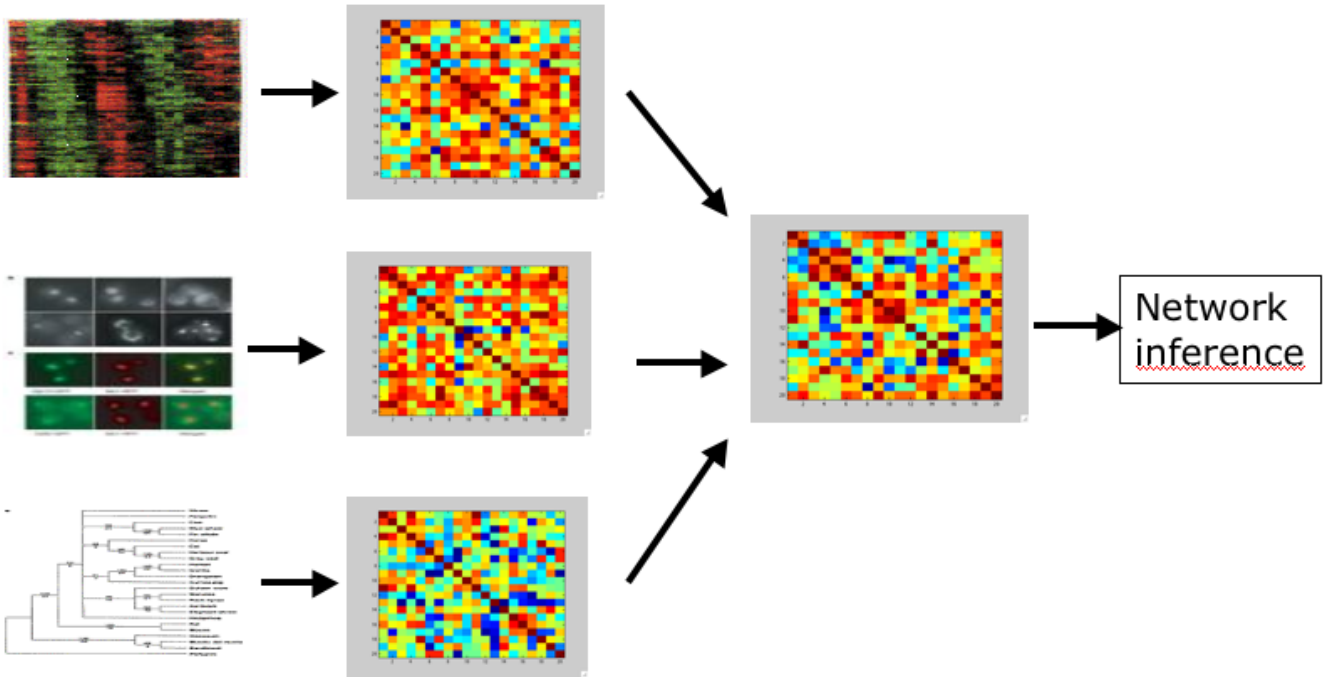
Part of the graph is known in advance

- Prior knowledge in model-based approaches
- Statistical / Machine learning approaches: learn from the known subnetwork a rule that can predict edges from genomic data

Genomic Data

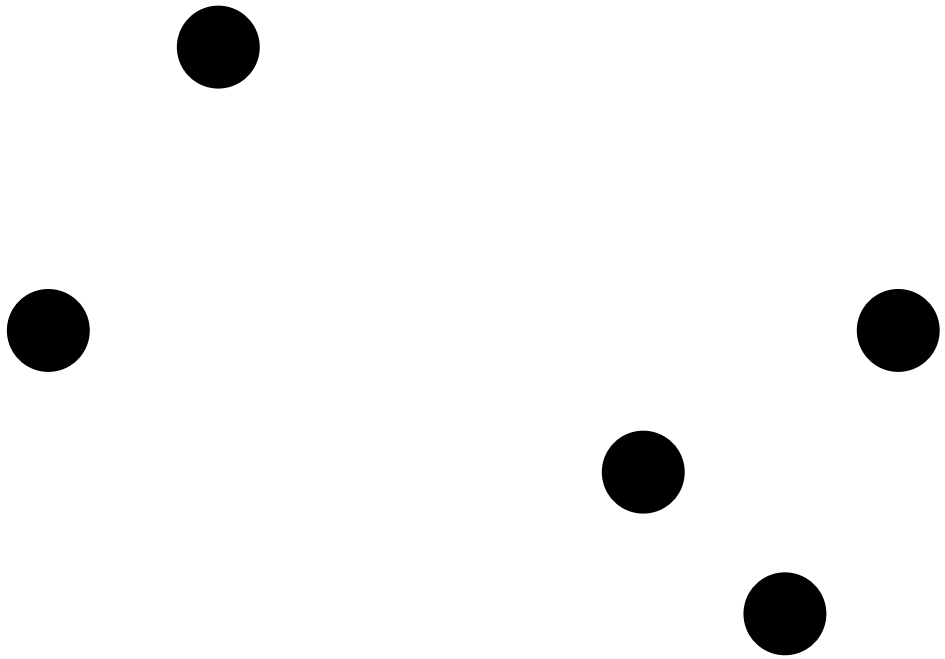
Data representation a distances

- We assume that each type of data (expression, sequences...) defines a (*negative definite*) distance between genes.
- Many such distances exist (cf kernel methods).
- Data integration is easily obtained by summing the distance to obtain an "integrated" distance



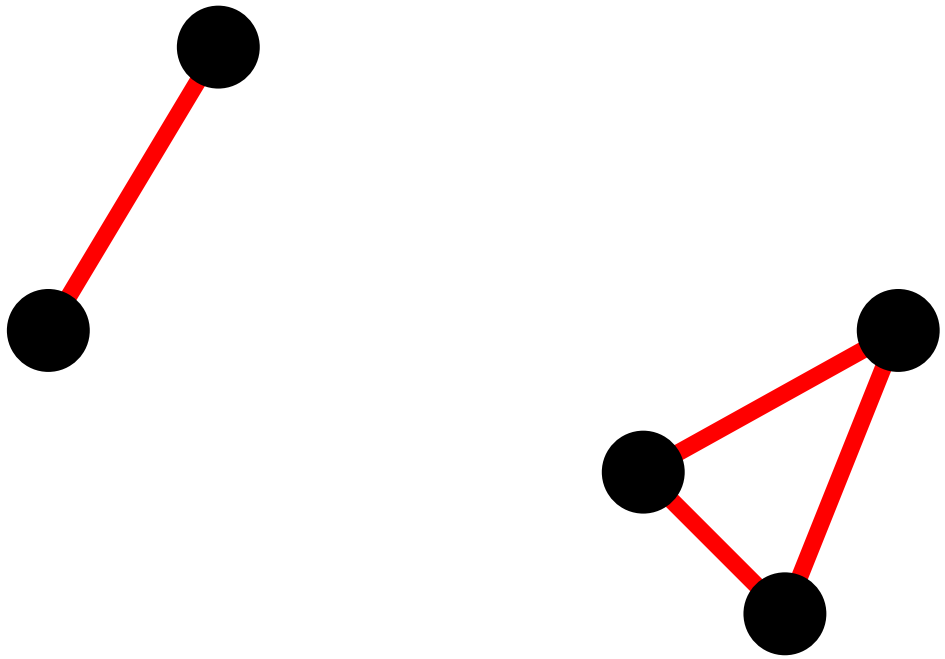
Method 1: Direct similarity-based prediction

- Motivation: "connect similar genes"
- Connect a and b if $d(a, b)$ is below a threshold.
- This is an unsupervised approach (no use of the known subnetwork).



Method 1: Direct similarity-based prediction

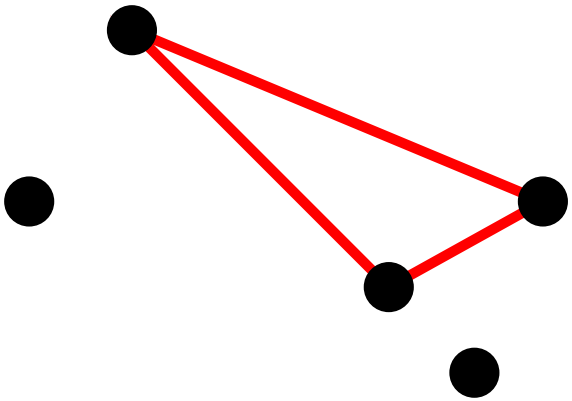
- Motivation: "connect similar genes"
- Connect a and b if $d(a, b)$ is below a threshold.
- This is an unsupervised approach (no use of the known subnetwork).



Method 2: metric learning

Metric learning

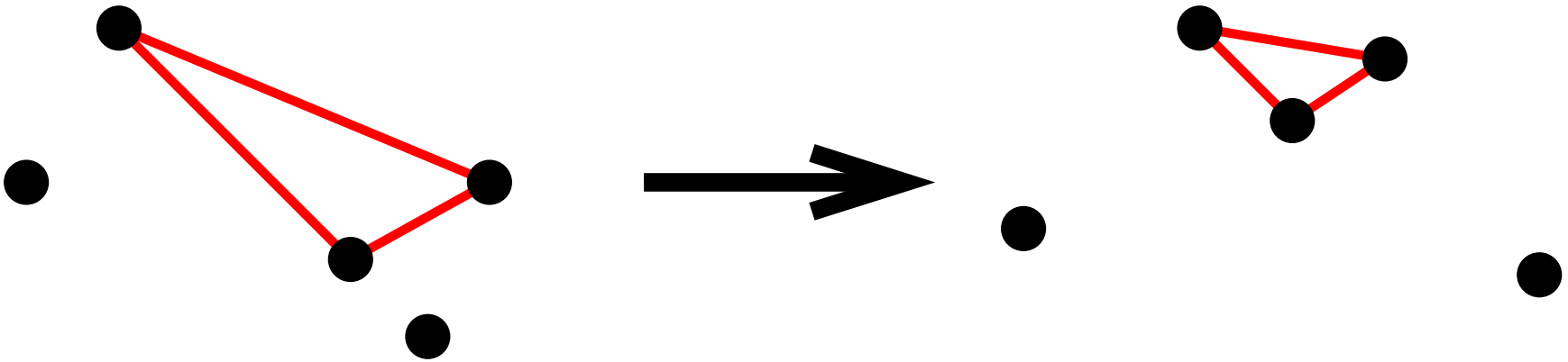
- Motivation: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Based on kernel CCA (Yamanishi et al., 2004) or kernel metric learning (V. and Yamanishi, 2005).



Method 2: metric learning

Metric learning

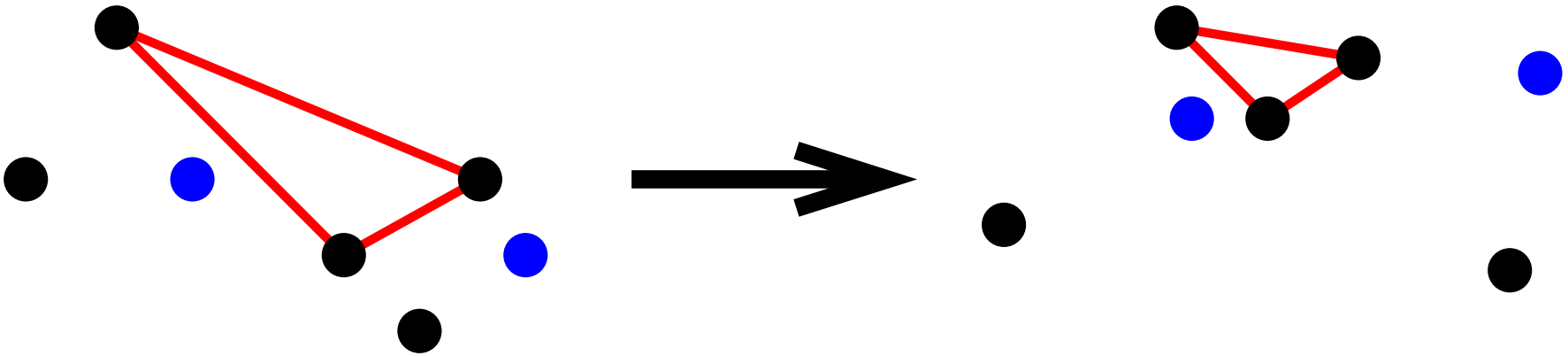
- Motivation: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Based on kernel CCA (Yamanishi et al., 2004) or kernel metric learning (V. and Yamanishi, 2005).



Method 2: metric learning

Metric learning

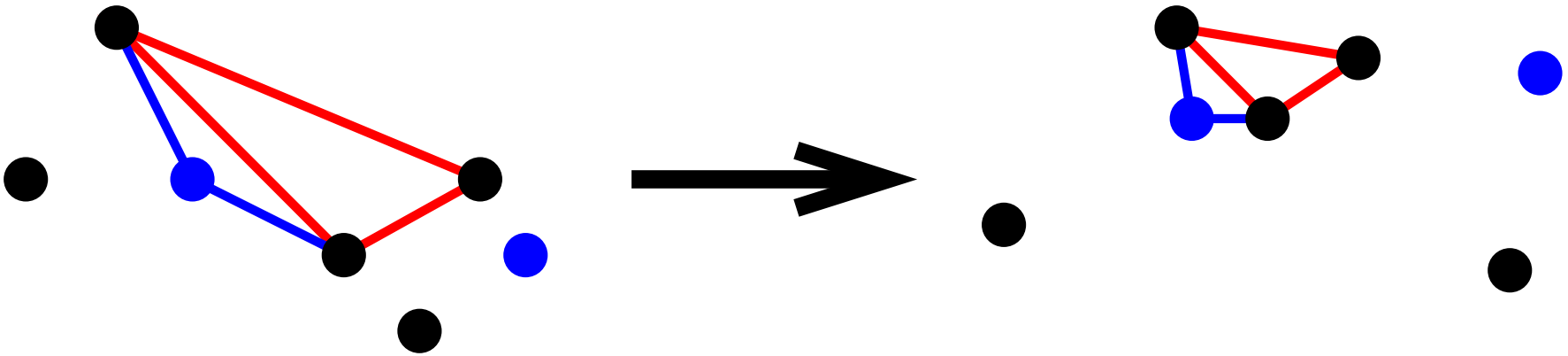
- Motivation: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Based on kernel CCA (Yamanishi et al., 2004) or kernel metric learning (V. and Yamanishi, 2005).



Method 2: metric learning

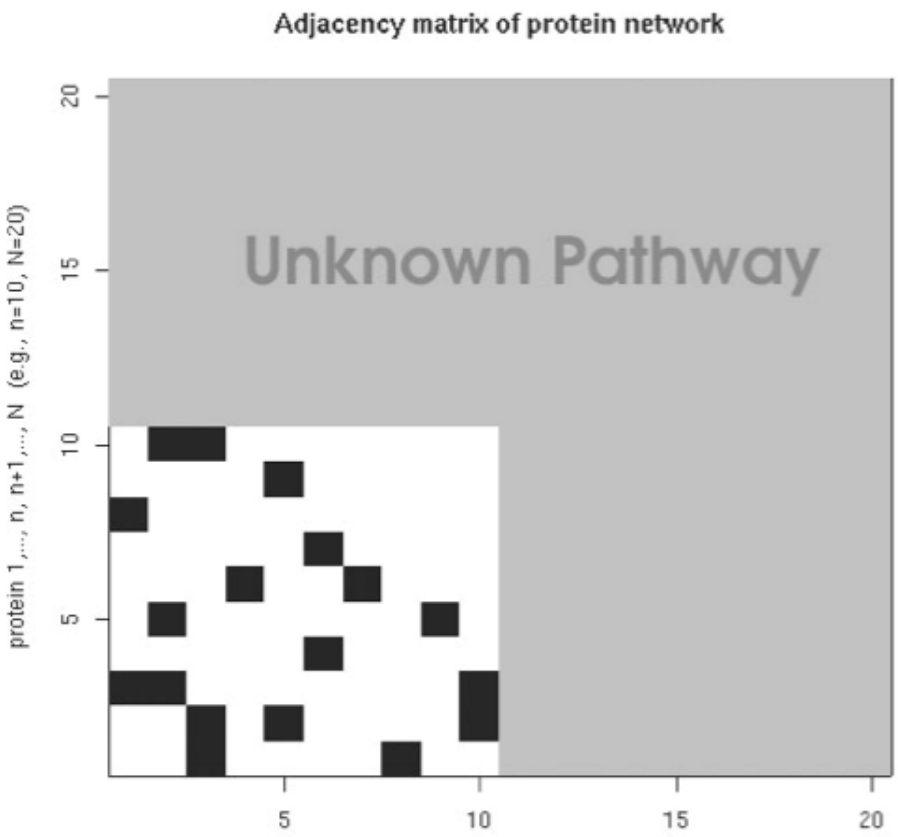
Metric learning

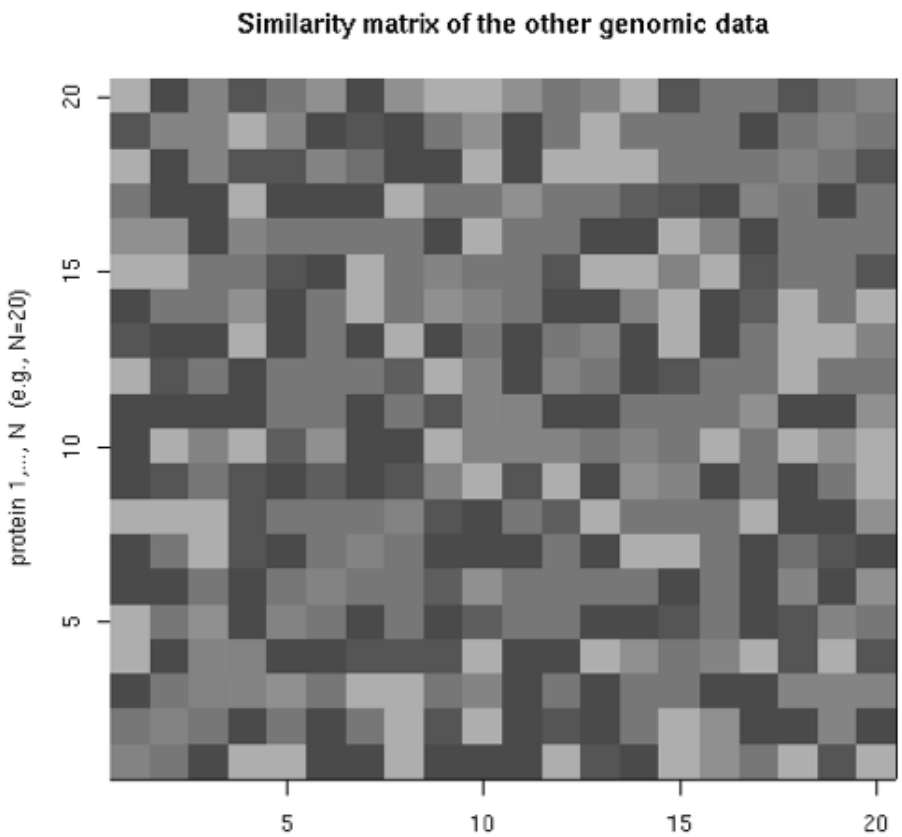
- Motivation: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Based on kernel CCA (Yamanishi et al., 2004) or kernel metric learning (V. and Yamanishi, 2005).



Method 3: Matrix completion

- Motivation: Fill missing entries in the adjacency matrix directly, by making it similar to (a variant of) the data matrix
- Method: EM algorithm based on information geometry of positive semidefinite matrices (Kato et al., 2005)





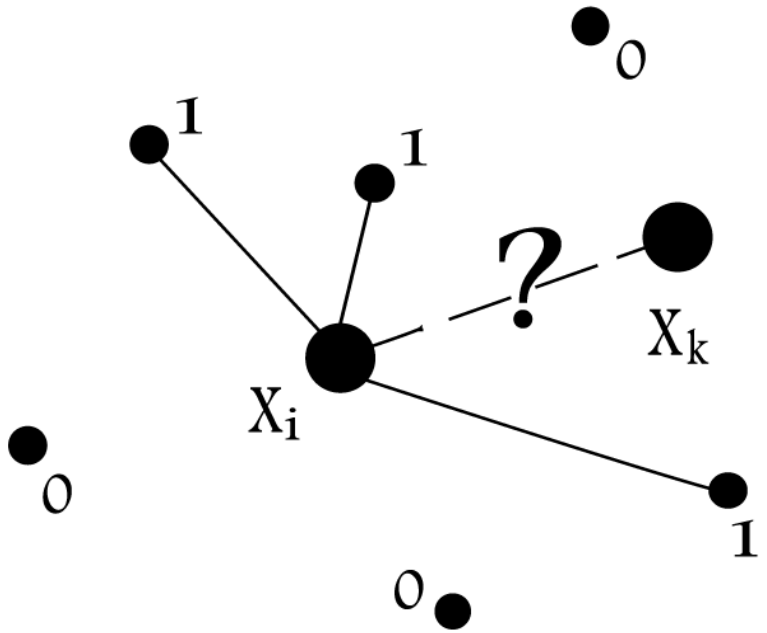
Method 4: Supervised binary classification

- A pair can be connected (1) or not connected (-1)
- Use known network as a training set for a SVM that will predict if new pair is connected or not
- Example: SVM with tensor product pairwise kernel (Ben-Hur and Noble, 2006):

$$K_{TTPK}\left((x_1, x_2), (x_3, x_4)\right) = K(x_1, x_3)K(x_2, x_4) + K(x_1, x_4)K(x_2, x_3) \ .$$

Method 5: Local predictions

- 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.



Experiments

Network

- Metabolic network (668 vertices, 2782 edges)
- Protein-protein interaction network (984 vertices, 2438 edges)

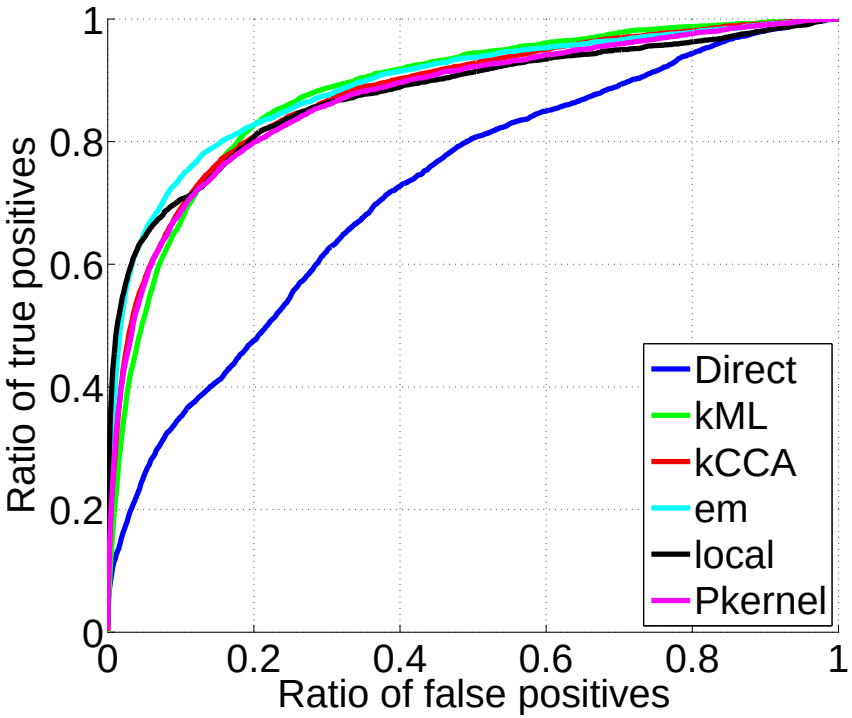
Data (yeast)

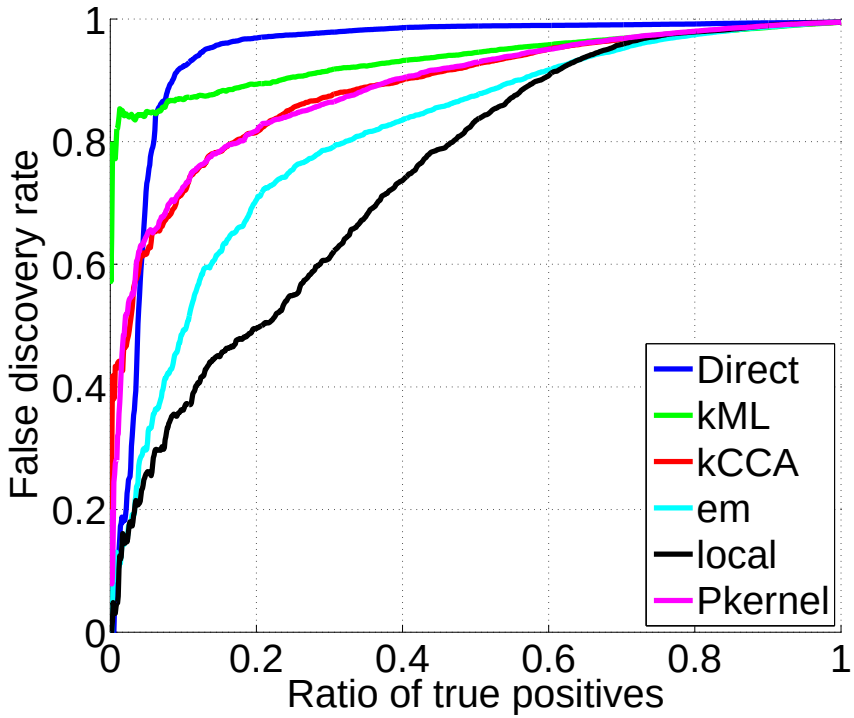
- Gene expression (157 experiments)
- Phylogenetic profile (145 organisms)
- Cellular localization (23 intracellular locations)
- Yeast two-hybrid data (2438 interactions among 984 proteins)

Method

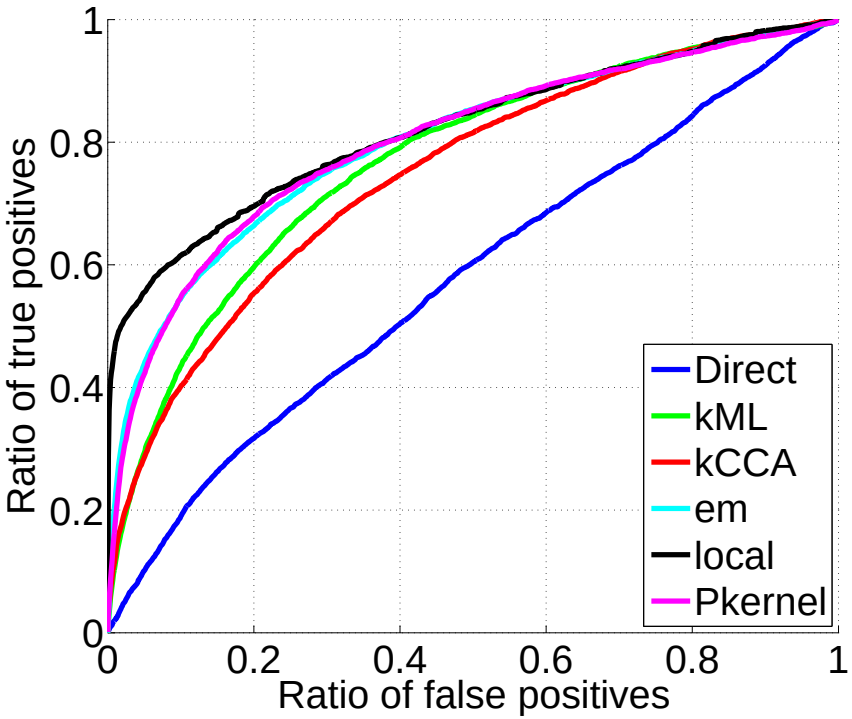
- 5-fold cross-validation
- Predict edges between test set and training set

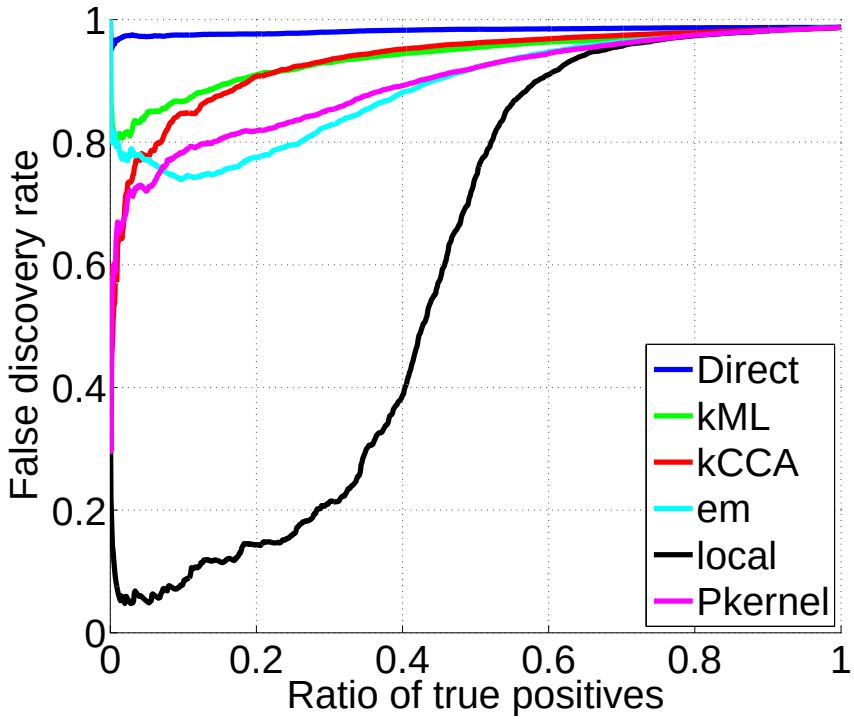
Results: protein-protein interaction



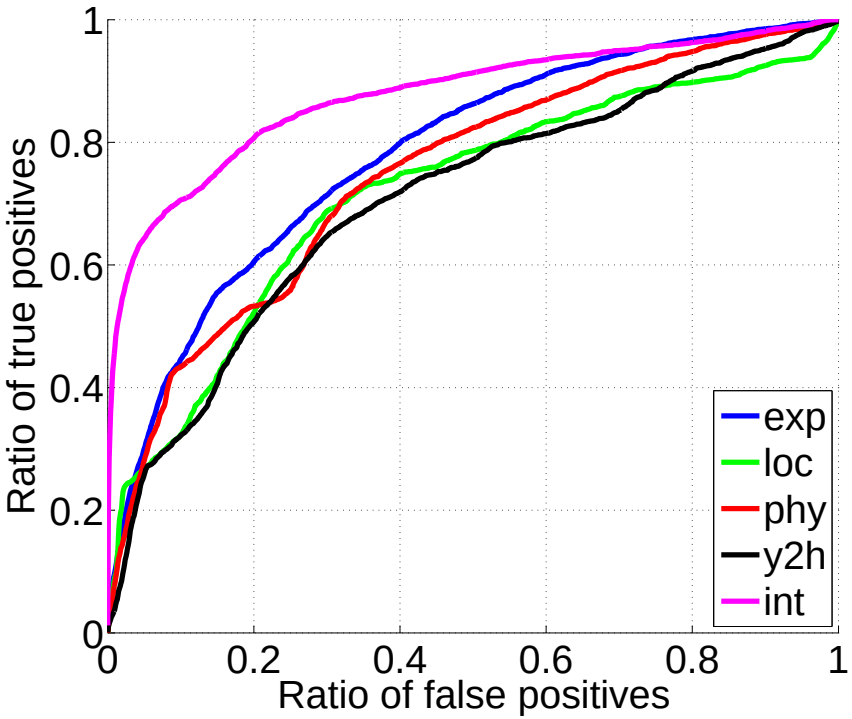


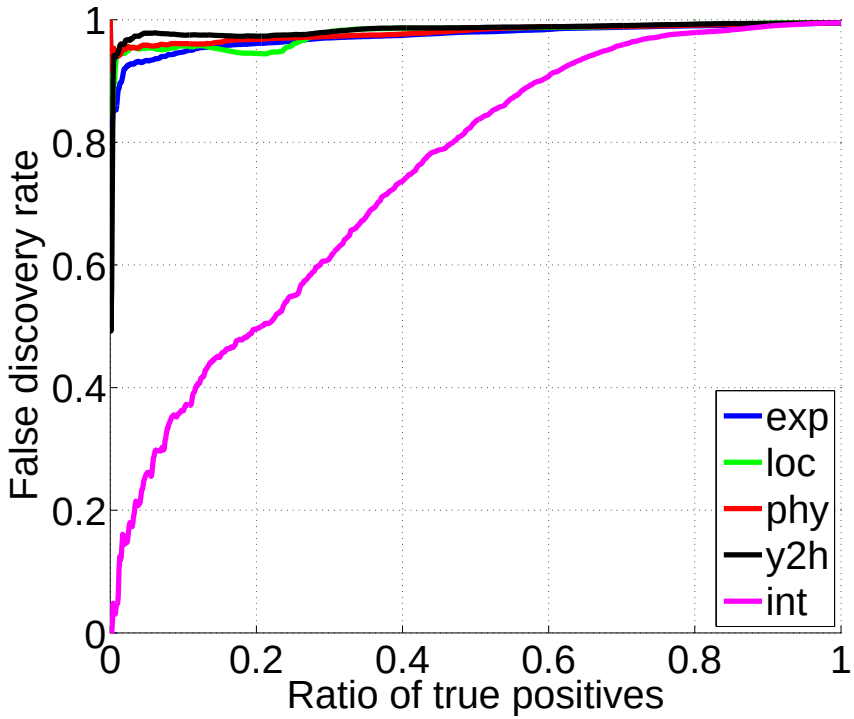
Results: metabolic gene network





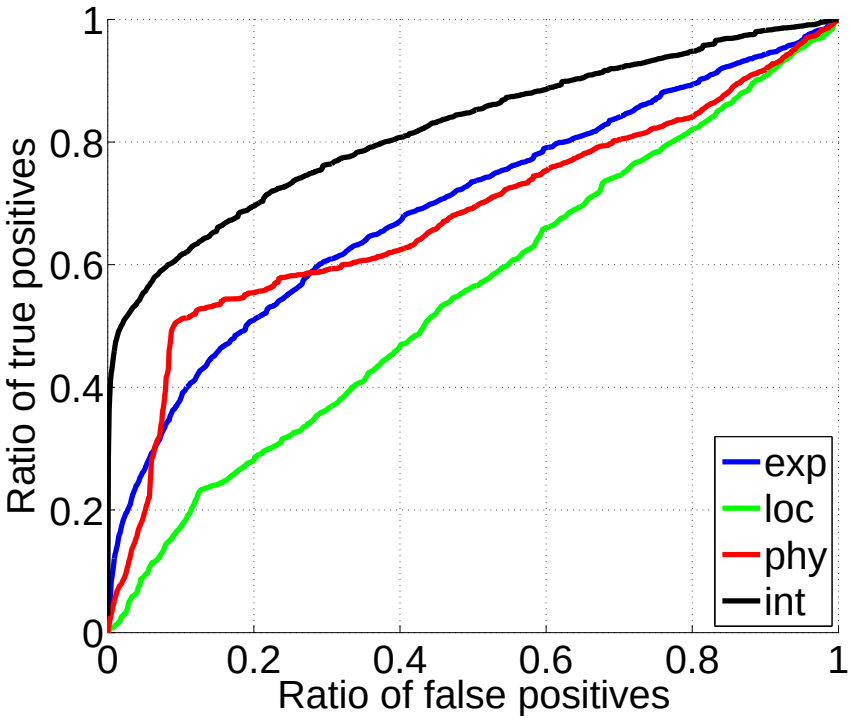
Results: effect of data integration

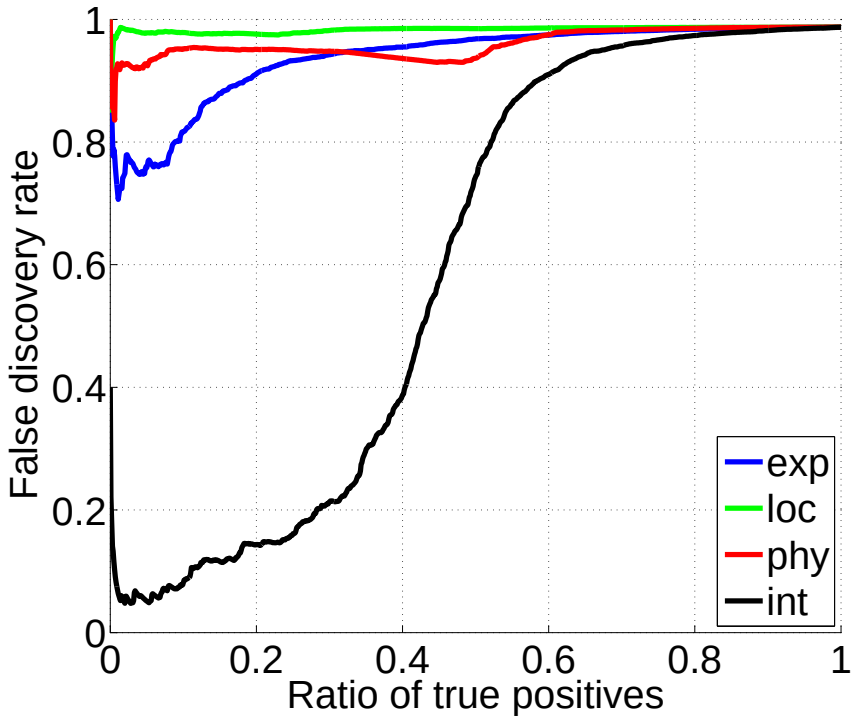




Local SVM, protein-protein interaction network.

Results: effect of data integration





Local SVM, metabolic gene network.

Conclusion

Summary

- A variety of methods have been investigated recently
- Some reach interesting performance on the benchmarks: Local SVM retrieve 45% of all true edges of the metabolic gene network at a FDR below 50%
- Valid for any network, but non-mechanistic model.
- Future work: experimental validation, improved data integration, semi-local approaches...

Acknowledgements

- Yoshihiro Yamanishi, Minoru Kanehisa (Univ. Kyoto): kCCA, kML
- Kevin Bleakley, Gerard Biau (Univ. Montpellier): local SVM