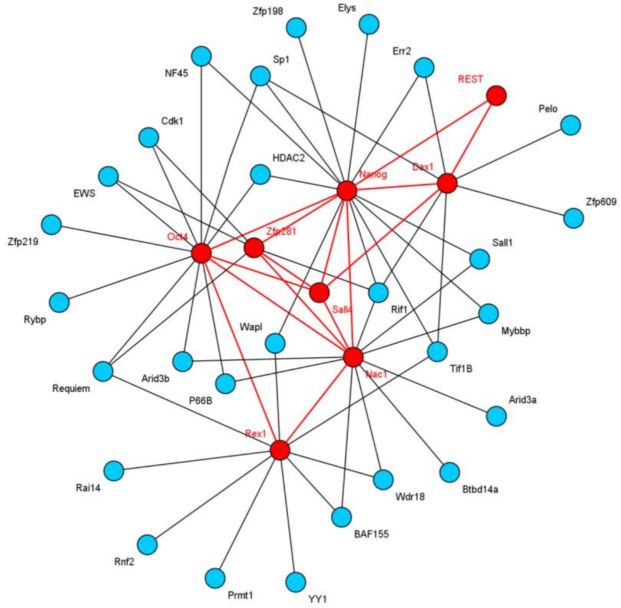
Big Data analytics for knowledge transfer among organisms while reconstructing Gene Regulatory Networks

Paolo Mignone (1,2), Gianvito Pio (1,2)
Domenica D'Elia (3), Michelangelo Ceci (1,2,4)

1 Department of Computer Science, University of Bari Aldo Moro, Bari 70125, Italy
2 National Interuniversity Consortium for Informatics (CINI), Roma 00185, Italy
3 CNR - Institute for Biomedical Technologies, Bari 70126, Italy
4 Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana 1000, Slovenia

Reconstruction of Gene Regulatory Networks



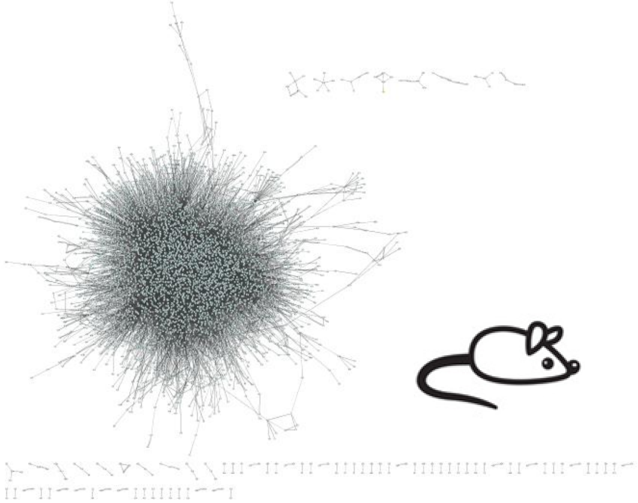
- A gene (or genetic) regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins
- When some control mechanisms are compromised, cells undergo a series of modifications that can bring to their transformation in **cancerous cells**

Motivations

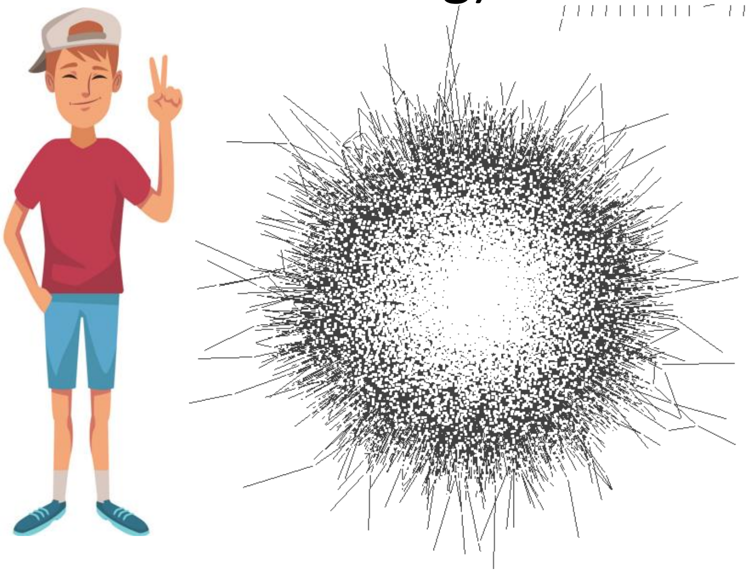
- Support biologists in the identification of new gene interactions for the better **understanding of diseases** related to gene network functions
- Existing methods suffer the limited number of **labelled examples** (i.e., validated interactions) and the absence of **negative examples** (i.e., confirmed absent interactions)

Goal

- Reconstructing the **Human Gene Network** using information from the **Mouse Gene Network** (via transfer learning)



Mouse «source domain»



Human «target domain»

The problem from a data mining perspective

- **Nodes**: genes
- **Edges**: gene interactions
- **Task**: gene network reconstruction via link prediction
- **Binary classification**:

✓ – the link exists – **POSITIVE LABEL**

✗ – the link does not exist – **NEGATIVE LABEL**

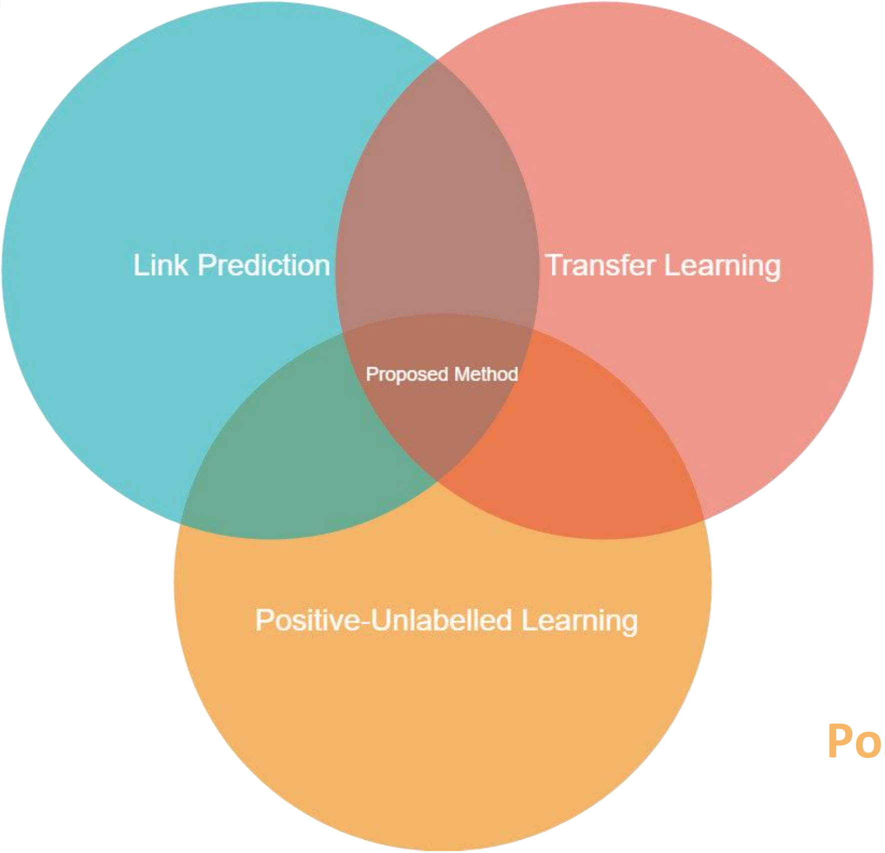
- **Training set**:

✓ – known existing links (ground truth)

? – no information about non-existing links

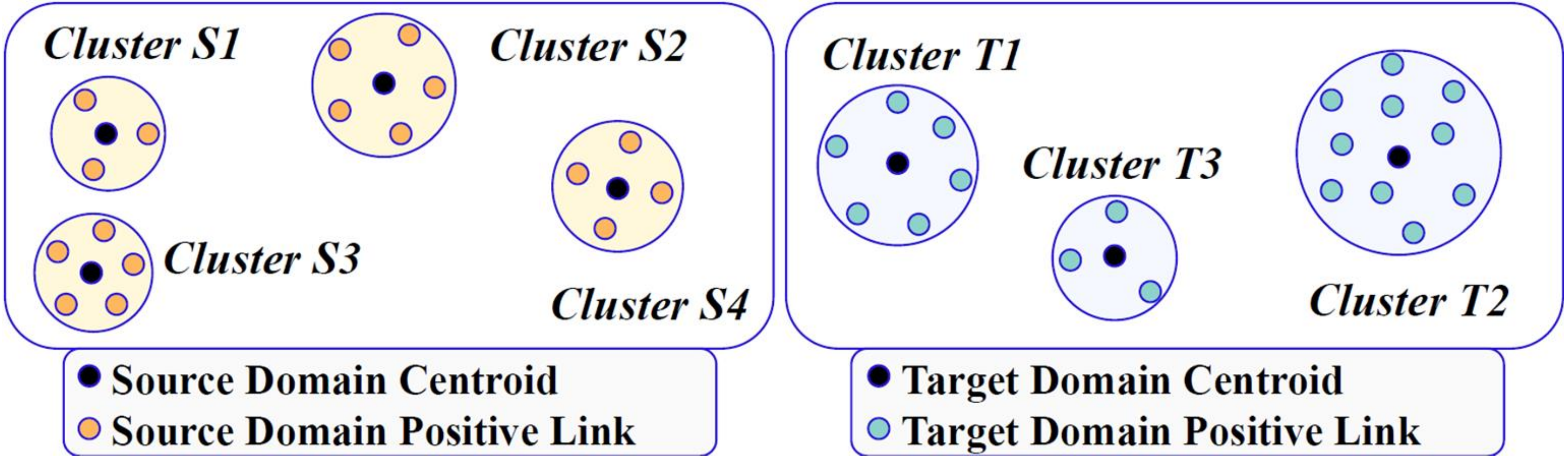
– positive-unlabelled learning setting

Proposed Method



Gene Network Reconstruction
solved as a
Link Prediction Task
supported by
Transfer Learning
in a
Positive-Unlabelled Learning Setting

Stage 1 – Clustering

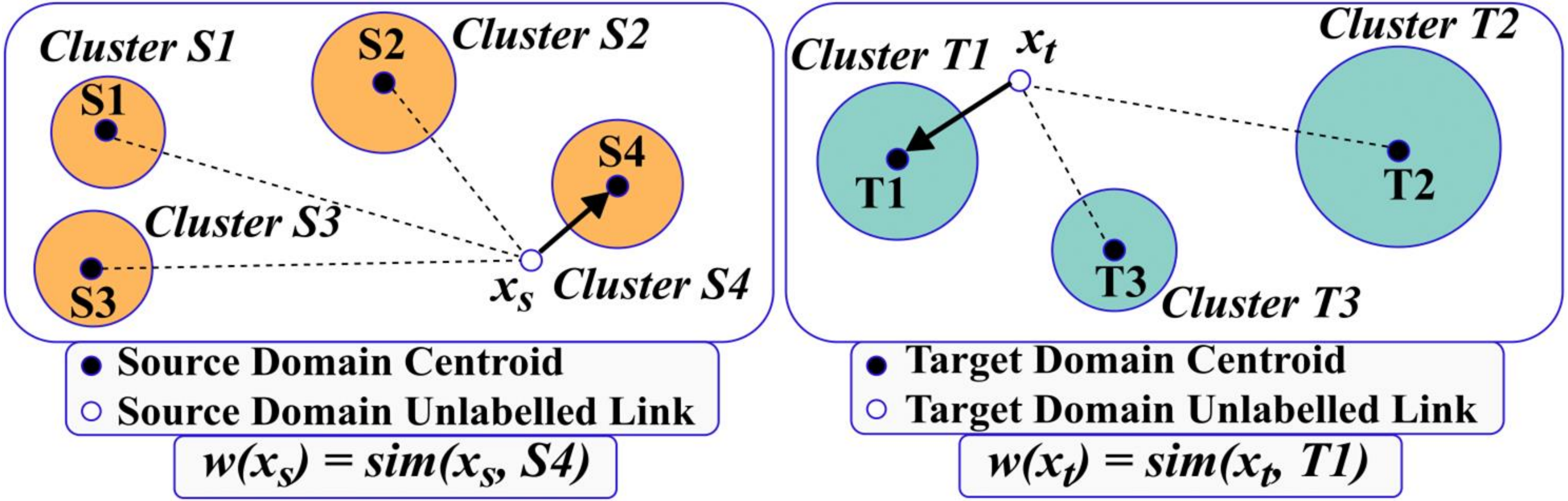


- **Clustering algorithm** on positive (i.e., validated) interactions on both the domains, separately, to obtain *n* and *m* clusters respectively

Stage 1 – Why clustering?

- 1) to distinguish the possible various concepts underlying the notion of **positive interaction**
- 2) to summarize the concept of positive interactions to simplify the weighting phase
  - computational **effort reduced** when computing distances between unlabelled instances and centroids w.r.t. to compare them to all the positive examples

Stage 2 – Weighting



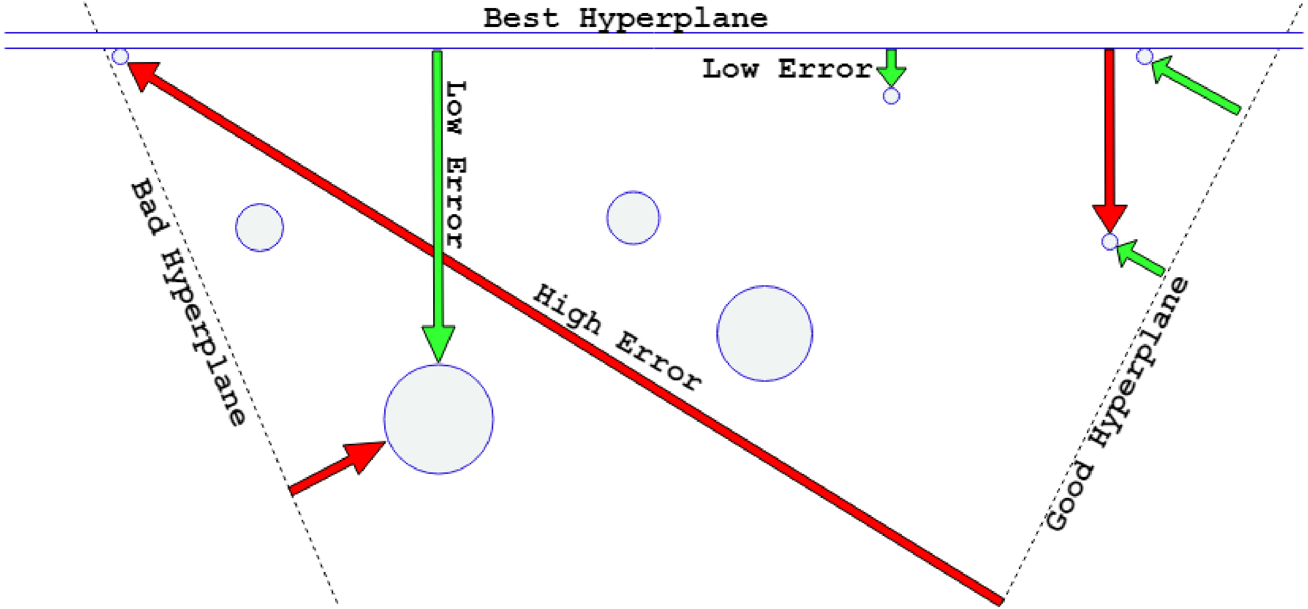
$w(x_s) = sim(x_s, S4)$

$w(x_t) = sim(x_t, T1)$

- Weighting of unlabelled interactions of the ***source domain S*** and the ***target domain T*** according to their distance with respect to the centroids

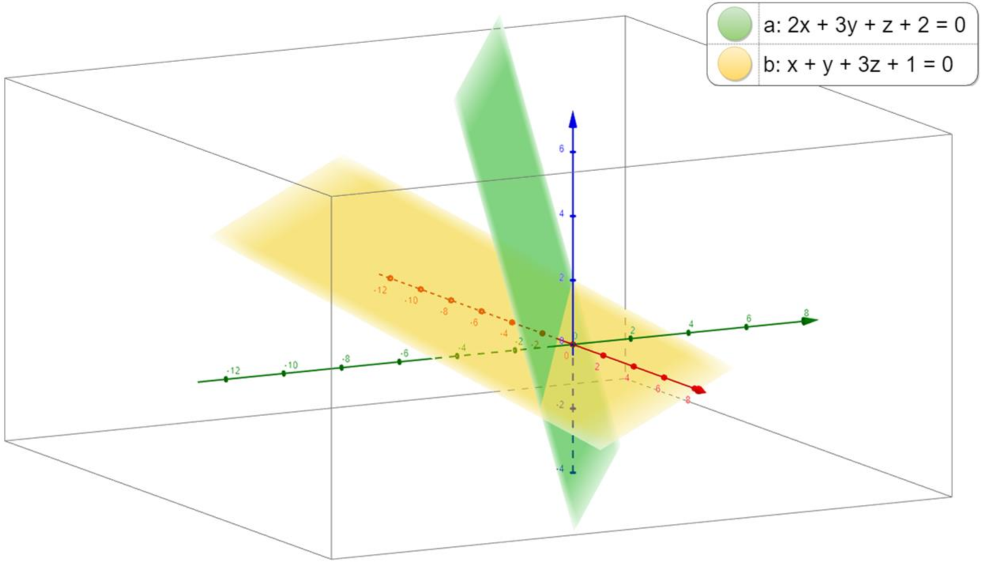
Stage 3 – Training

- Training of a classifier (SVM-based) that is **able to exploit weights** on the instances: **Weighted SVM (WSVM)**



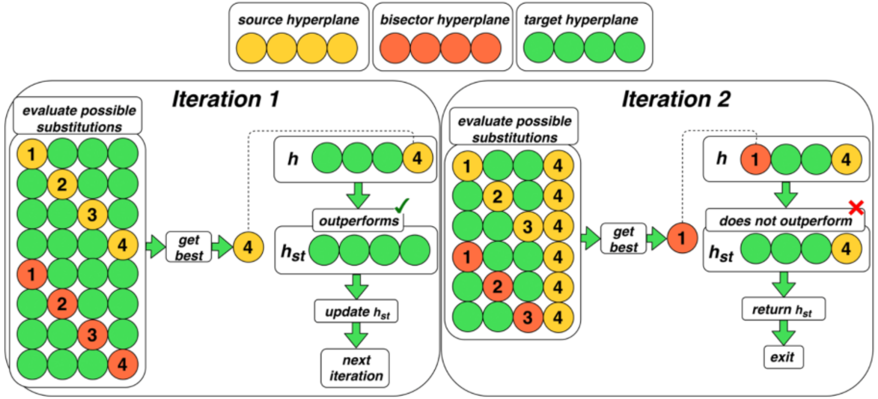
Stage 3 – Training

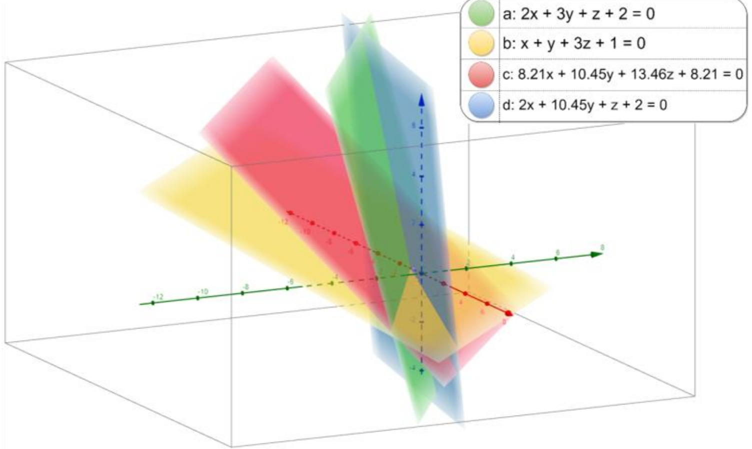
- Training of two different WSVM classifiers for the source and the target domains separately
- **WSVM source domain model**
- **WSVM target domain model**



Stage 4 – Model Combination

- Compute the **bisector hyperplane**
- Perform a hyperplane combination based on data driven coefficient substitutions
- **Input:**
  - **source coefficients**
  - **target coefficients**
  - **bisector coefficients**
- **Output:**
  - **new combined hyperplane**





Experiments

- Gene interactions dataset

	Positive*	Unlabelled**
Mouse (source)	14613	235706
Human (target)	235706	235706

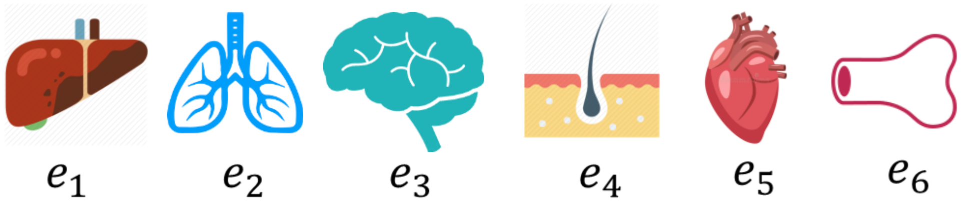
*BIOGRID - https://thebiogrid.org/

** Gene Expression Omnibus - https://www.ncbi.nlm.nih.gov/geo/

Experiments

- Gene representation

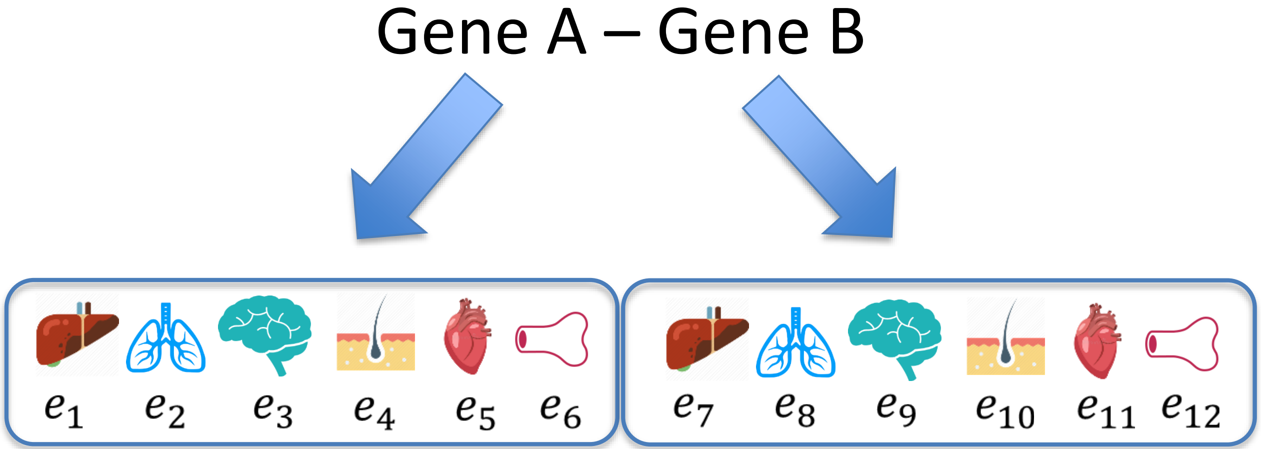
Features: average gene expression levels measured for specific tissues in control samples



$$e_i \in [0,1]$$

Experiments

- Gene interaction representation



Gene A features

Gene B features

$$e_i \in [0,1]$$

Experiments

We compared our method, indicated as **BioSfer** in the results, with two baselines:

- **no_transfer:** WSVM with Platt scaling, learned only from the target network (i.e., from the human gene network). Allows us to evaluate the **contribution of the source domain**

- **union:** WSVM with Platt scaling, learned from a single dataset consisting of the union of the instances coming from both mouse and human. Allows us to evaluate the **effect of our weighting strategy**

Experiments

Transfer Learning Competitors

- **JGSA** hang, J., Li, W., and Ogunbona, P. (2017). Joint geometrical and statistical alignment for visual domain adaptation. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.
- **BDA** Wang, J., Chen, Y., Hao, S., Feng, W., and Shen, Z. (2017). Balanced distribution adaptation for transfer learning. In ICDM 2017
- **TJM** Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2014). Transfer joint matching for unsupervised domain adaptation. In CVPR 2014

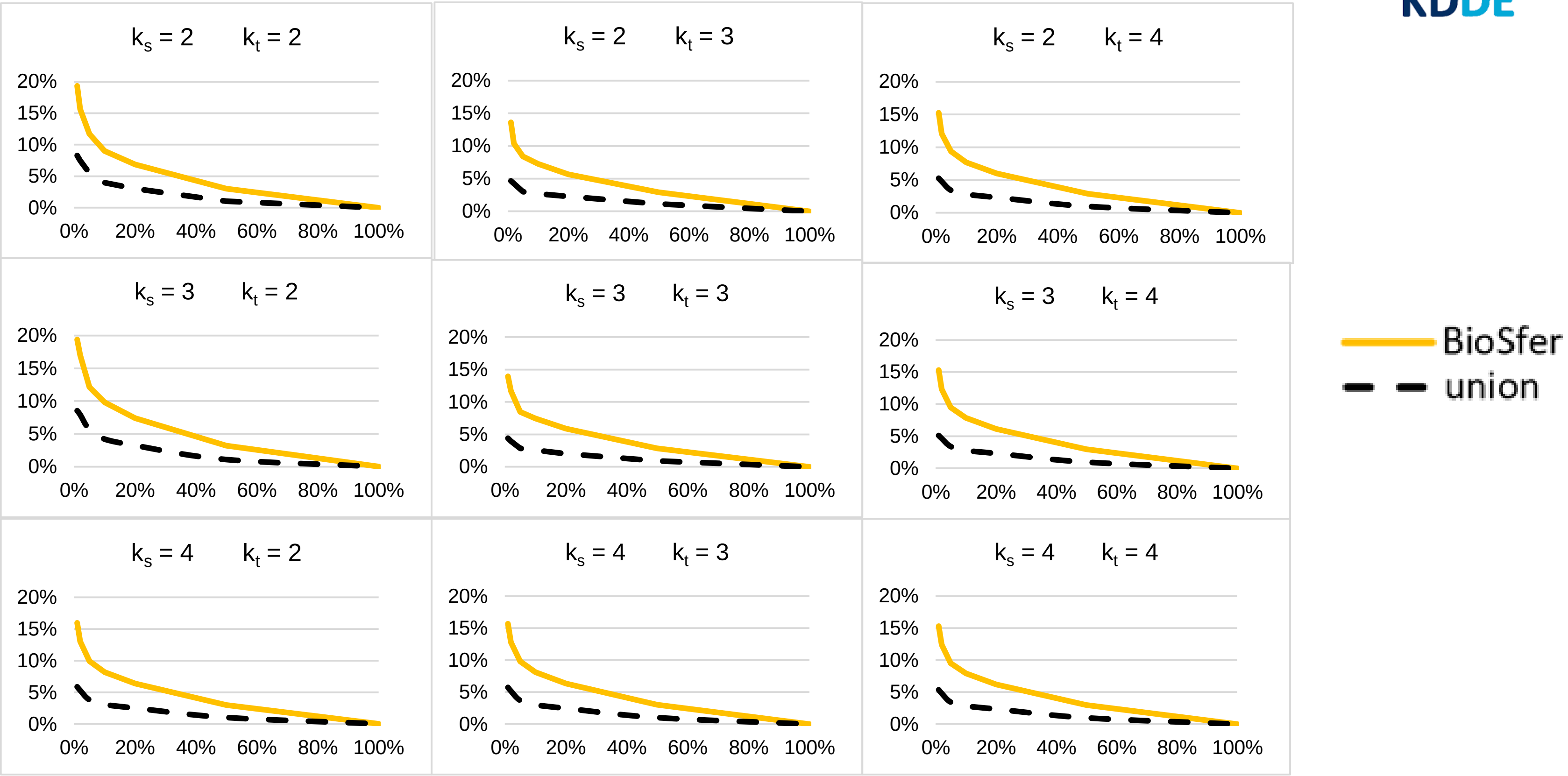
Gene Regulatory Network Reconstruction Competitor

- **GENIE3** Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, Pierre Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. In PLOS ONE September 28, 2010, https://doi.org/10.1371/journal.pone.0012776

Experimental Setting

- Evaluation measures:
  - Recall@k
  - Area Under Recall@K curve (AUR@K)
  - Area Under ROC curve (AUROC)
  - Area Under Precision Recall curve (AUPR)
- Clustering setting: $k_s, k_t \in \{2, 3, 4\}$
- 10 fold cross-validation

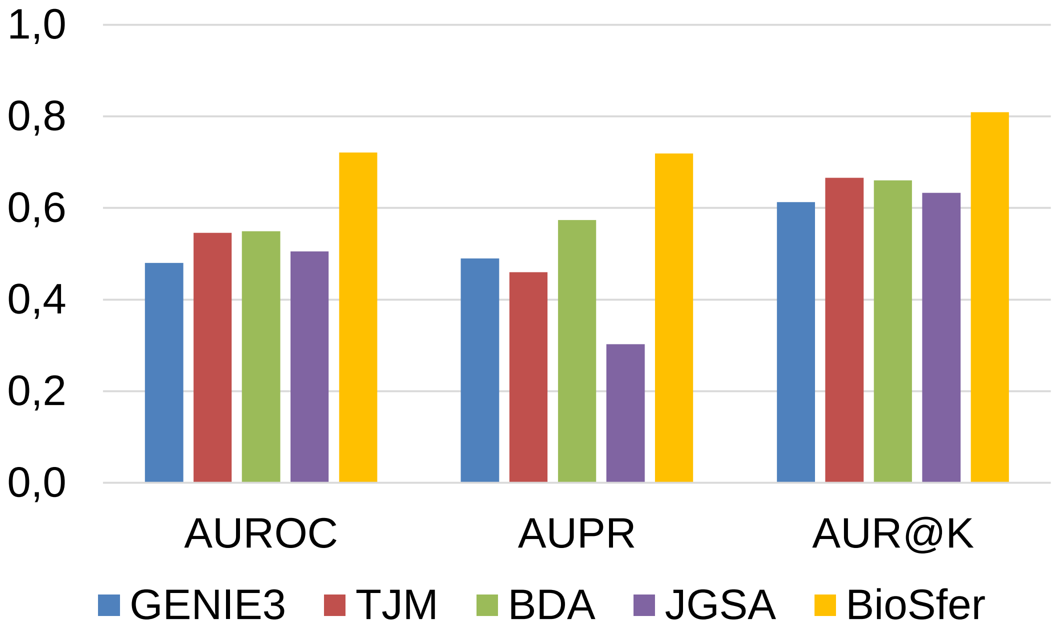
Results


$k_s = 2$ $k_t = 2$
$k_s = 2$ $k_t = 3$
$k_s = 2$ $k_t = 4$
$k_s = 3$ $k_t = 2$
$k_s = 3$ $k_t = 3$
$k_s = 3$ $k_t = 4$
$k_s = 4$ $k_t = 2$
$k_s = 4$ $k_t = 3$
$k_s = 4$ $k_t = 4$
20%
15%
10%
5%
0%
0%
20%
40%
60%
80%
100%
BioSfer
union

Results

KDDE

Comparison between **BioSfer** and competitors with the **best configuration**



Qualitative Analysis

- Top-1 Ranked Predicted Interaction: **NBPF8P - ND4**, score: 0.948
- Top-2 Ranked Predicted Interaction: **LINC00657 - RPL39**, score: 0.945
- Not identified by **g:Profiler** (*https://biit.cs.ut.ee/gprofiler/gost*)

NBPF8P - ND4

Since the correlation between the **impaired mitochondrial respiratory chain** function and the pathogenesis of several **neurological diseases** is well-known [A], the prediction provided by BioSfer is **biologically reasonable**

LINC00657 - RPL3

The relationship between these two genes is plausible because of the recent discovery of the existence of **ribosome-associated non-coding RNAs (rancRNAs).** Ribosomes can be the target for numerous small and long non-coding RNAs in various organisms [B]

[A] R. K. Chaturvedi and M. F. Beal. ***Mitochondrial diseases of the brain.*** Free Radical Biology and Medicine, 63:1–29, oct 2013.

[B] A. Pircher, J. Gebetsberger, and N. Polacek***. Ribosome-associated ncRNAs: An emerging class of translation regulators.*** RNA Biology, 11(11):1335–1339, nov 2014.

Conclusion

- The knowledge about the mouse gene network is helpful to **better reconstruct** the human gene network
- BioSfer is able to exploit the information of unlabelled gene pairs in order to better identify a set of **existing gene interactions**

Future Works

- **Multiple source networks** for the reconstruction of the target network

Thanks for your attention