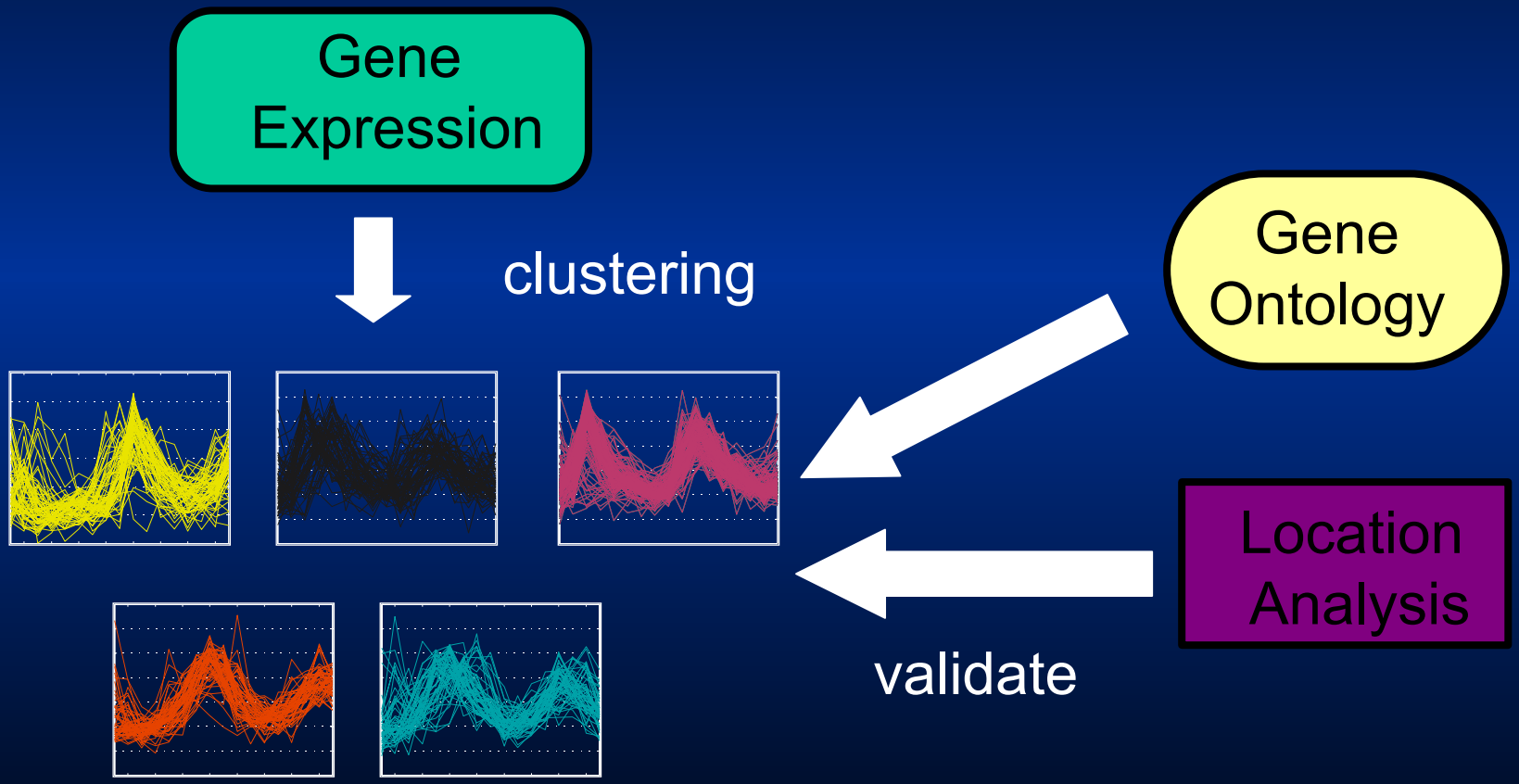
On the Feasibility of Heterogeneous Analysis of Large Scale Biological Data

Ivan G. Costa Filho
Alexander Schliep

Computational Biology Department
Max-Planck-Institute for Molecular Genetics, Berlin

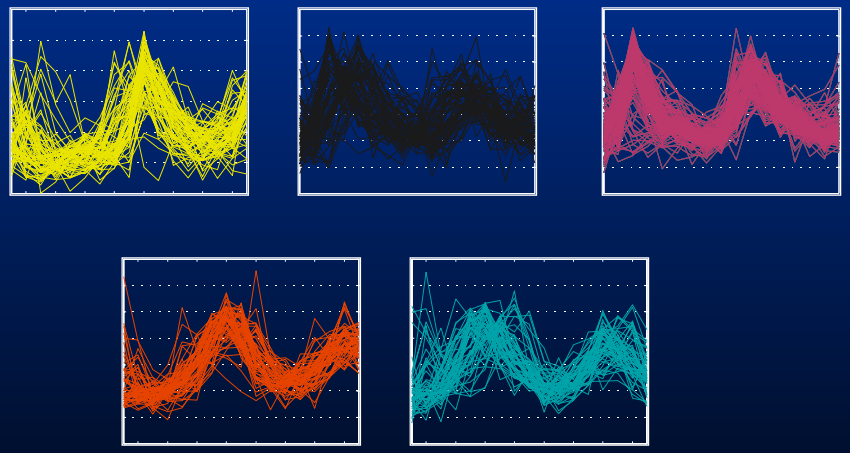
Motivation


Gene Expression
clustering
Gene Ontology
Location Analysis
validate

Motivation

Gene Expression

‘heterogeneous’ clustering



Gene Ontology

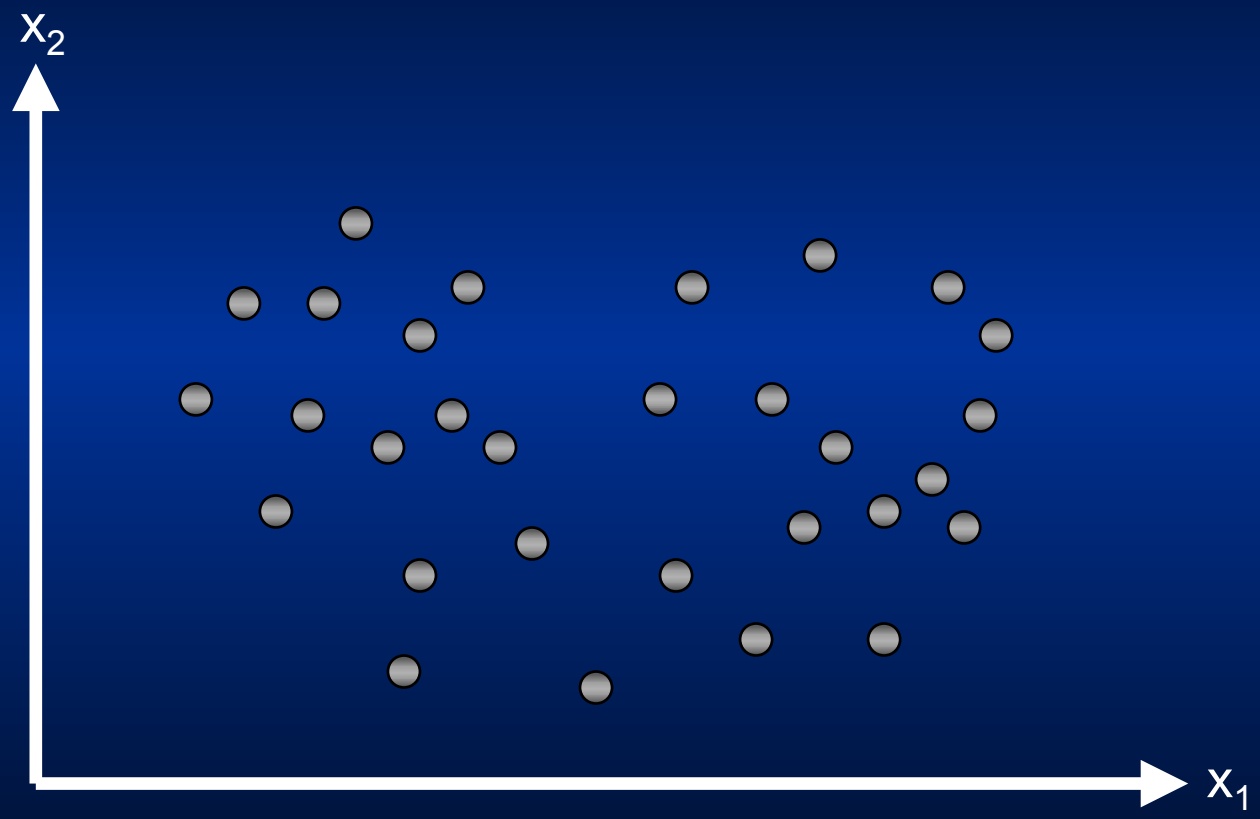
Location Analysis

Our Approach

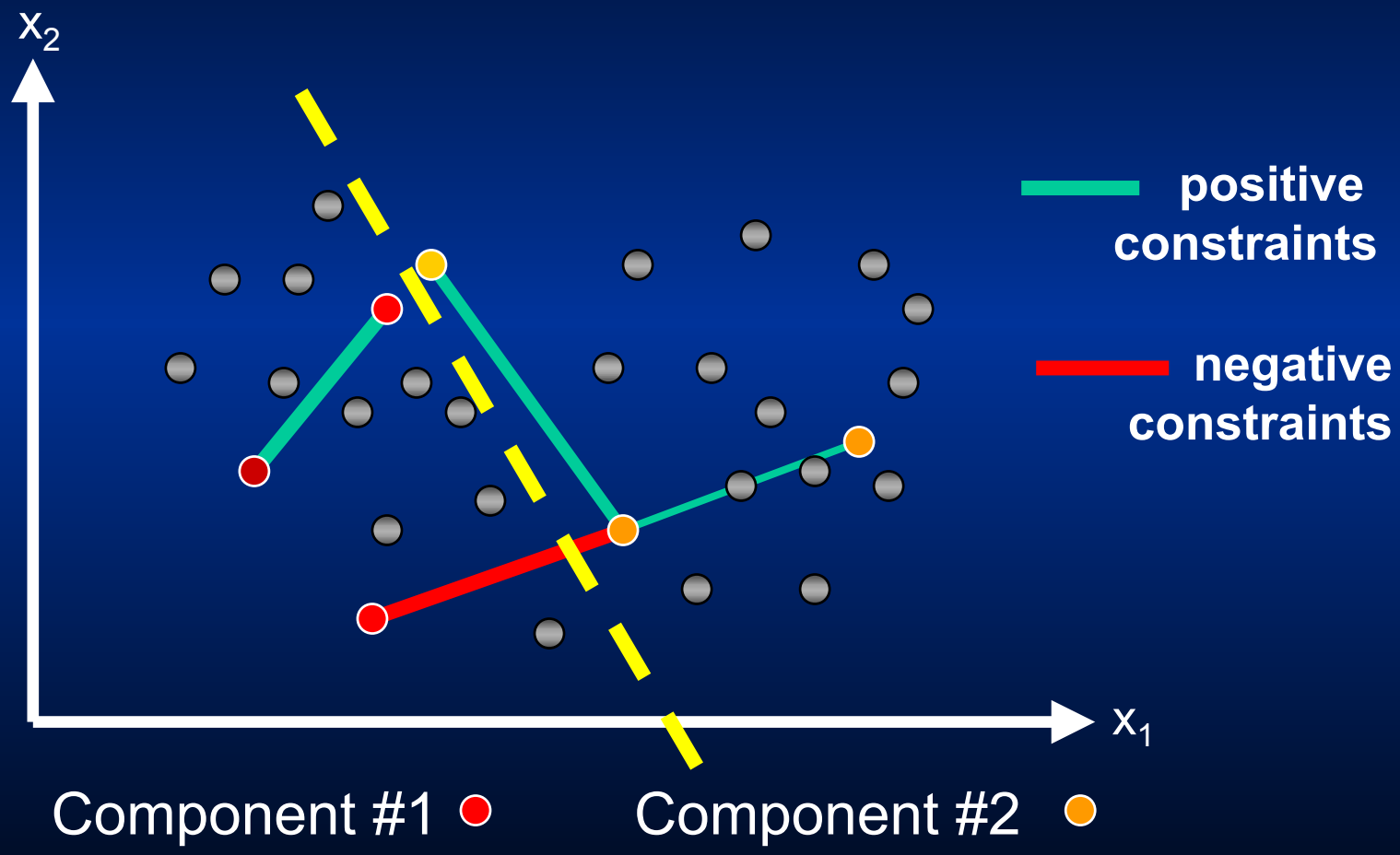
Semi-supervised learning

- encode location analysis and GO as soft pair-wise constraints
- clustering with constraints (Lu and Leen, 2005, Lange *et al.*, 2005)

Clustering with Constraints



Clustering with Constraints



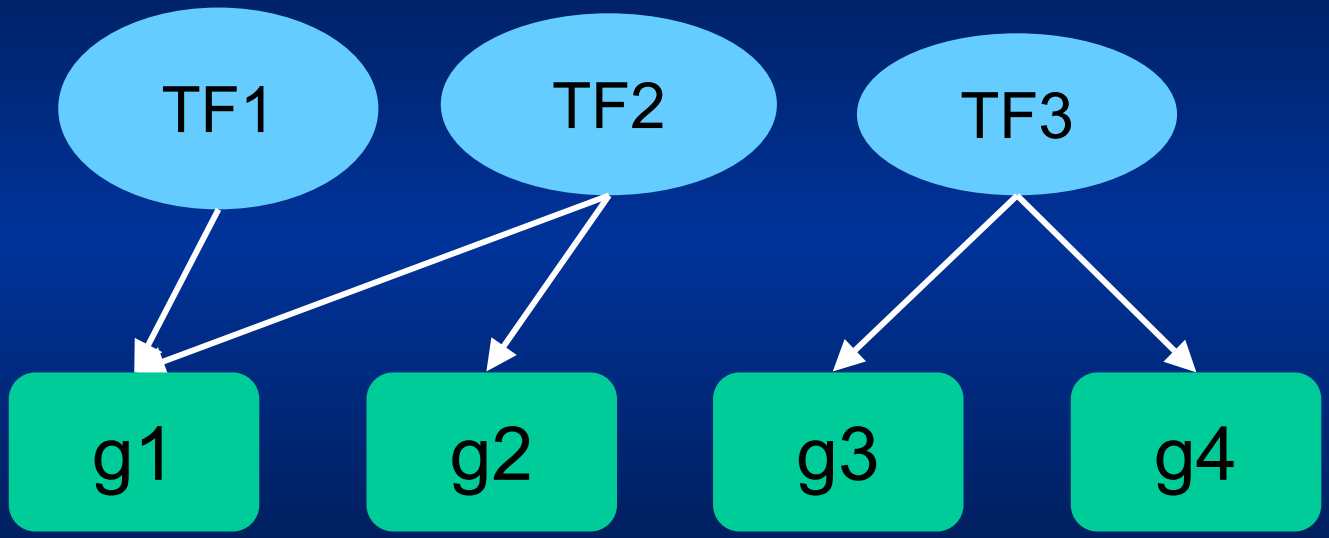
Clustering with Constraints

Idea:penalize the complete log likelihood

$$\exp \sum_i \sum_{j \neq i} -w_{ij}^{+} 1\{y_i \neq y_j\} \lambda^{+}$$

where $\boldsymbol{Y}$ are the cluster assignments and W^{+} the positive pair-wise constraints [Lange *et al.,* 2005].

Soft pair-wise Constraints Location Analysis



$w^{+}_{(1,2)} = 0.5$ $w^{+}_{(2,3)} = 0.0$ $w^{+}_{(3,4)} = 1.0$

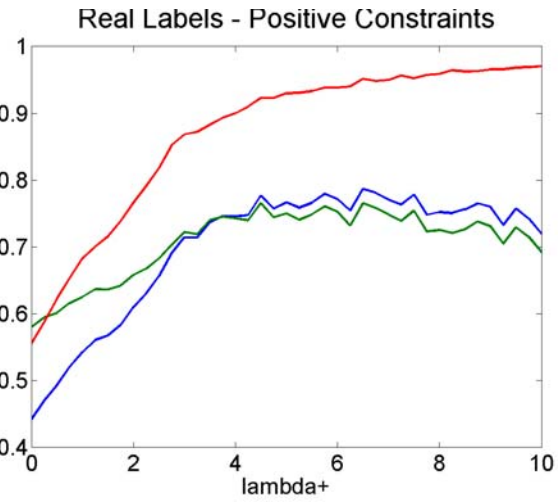
Experiments

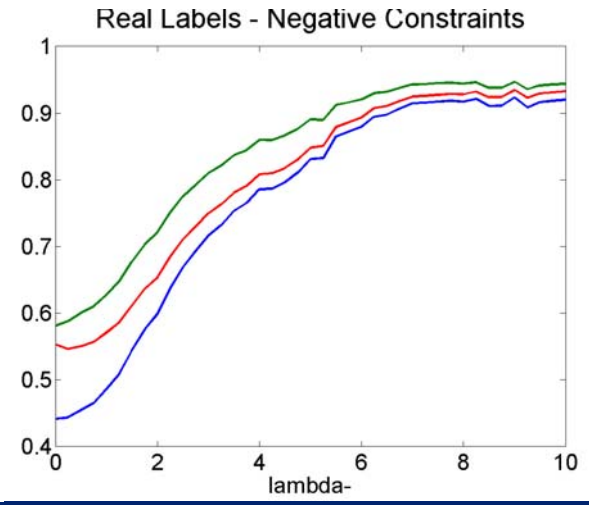
Data

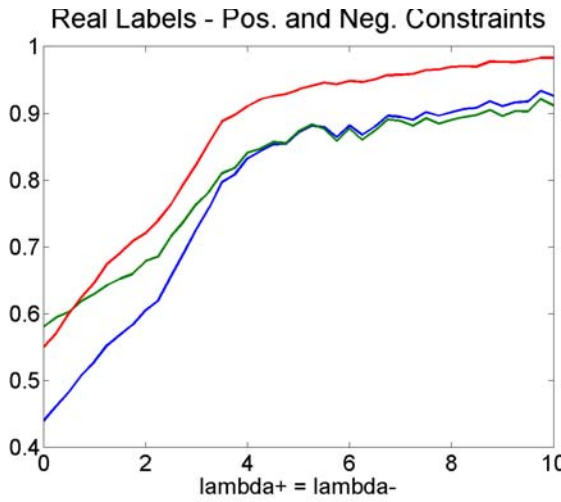
- Gene expression data
  - time-courses of 384 genes during mitotic cell division in Yeast (Cho, 1998)
  - expert classification into ‘five’ cell-cycle phases
  - modeled with diagonal multivariate Gaussian
- Constraints
  - true labels
  - transcription factor location analysis (Lee, 2002)
  - Gene Ontology (not shown)

Results

Constraints from True Labels





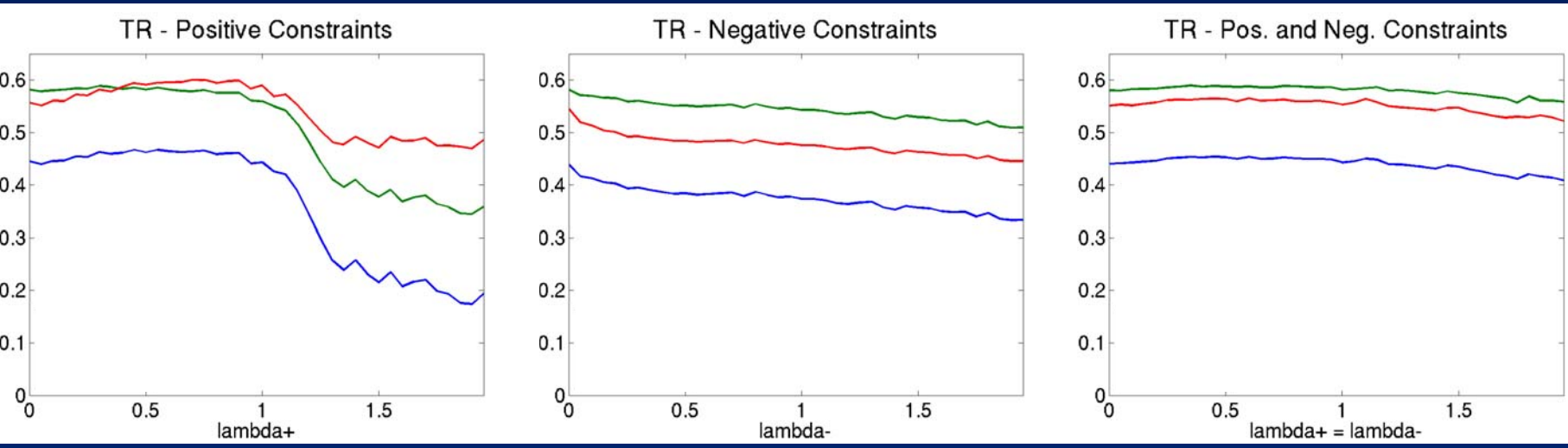


corrected Rand
specificity
sensitivity

5% of gene pairs constrained

λ^+ and λ^- - constraints weight

Constraints from Location Analysis



- corrected Rand
- specificity
- sensitivity

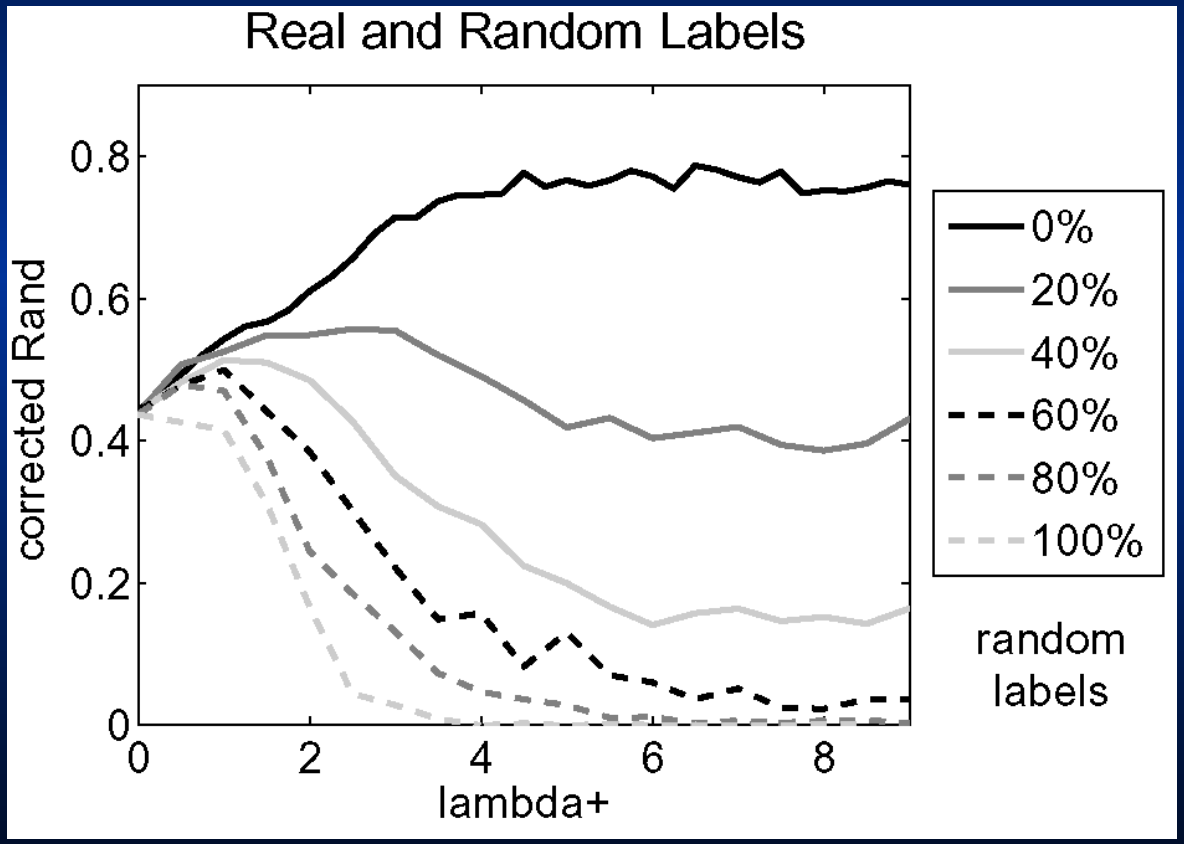
40% of gene pairs constrained

λ^+ and λ^- - constraints weight

Possible Explanations

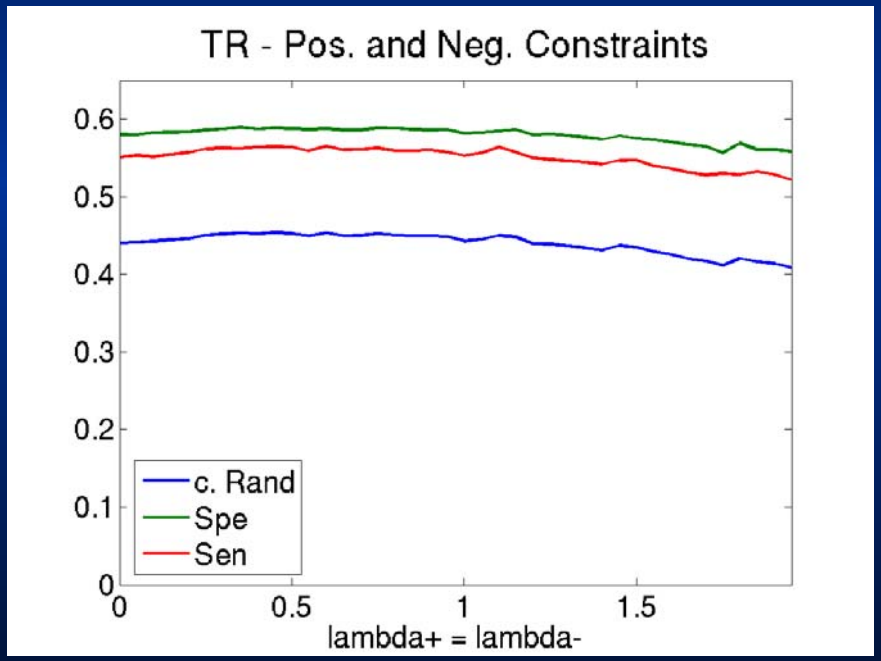
- Noise in the data
- Non-specific information content

Constraints from True and Random Labels

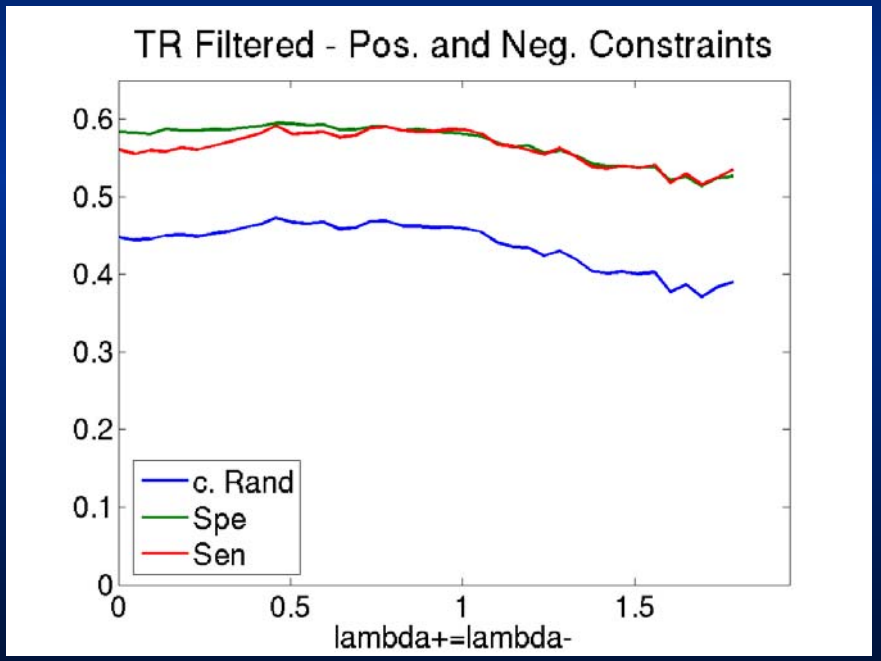


Filtered Constraints from Location Data

Non-filtered



Filtered



Conclusions

- Method works with few ‘true’ constraints
  - <1% high quality biological annotation yield similar results [Schliep, 2004]
- Insignificant performance gain is obtained with biological constraints
  - high λ^+/ λ^- deteriorate greatly the results
- Enriched transcription factors (filtered Location Data) did not yield significant improvement

Thanks.

http://algorithmics.molgen.mpg.de