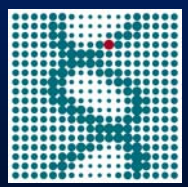
Mixture Model Estimation with Constraints:

Analysis of time-course gene expression with heterogeneous data

Ivan G. Costa Filho
Alexander Schliep

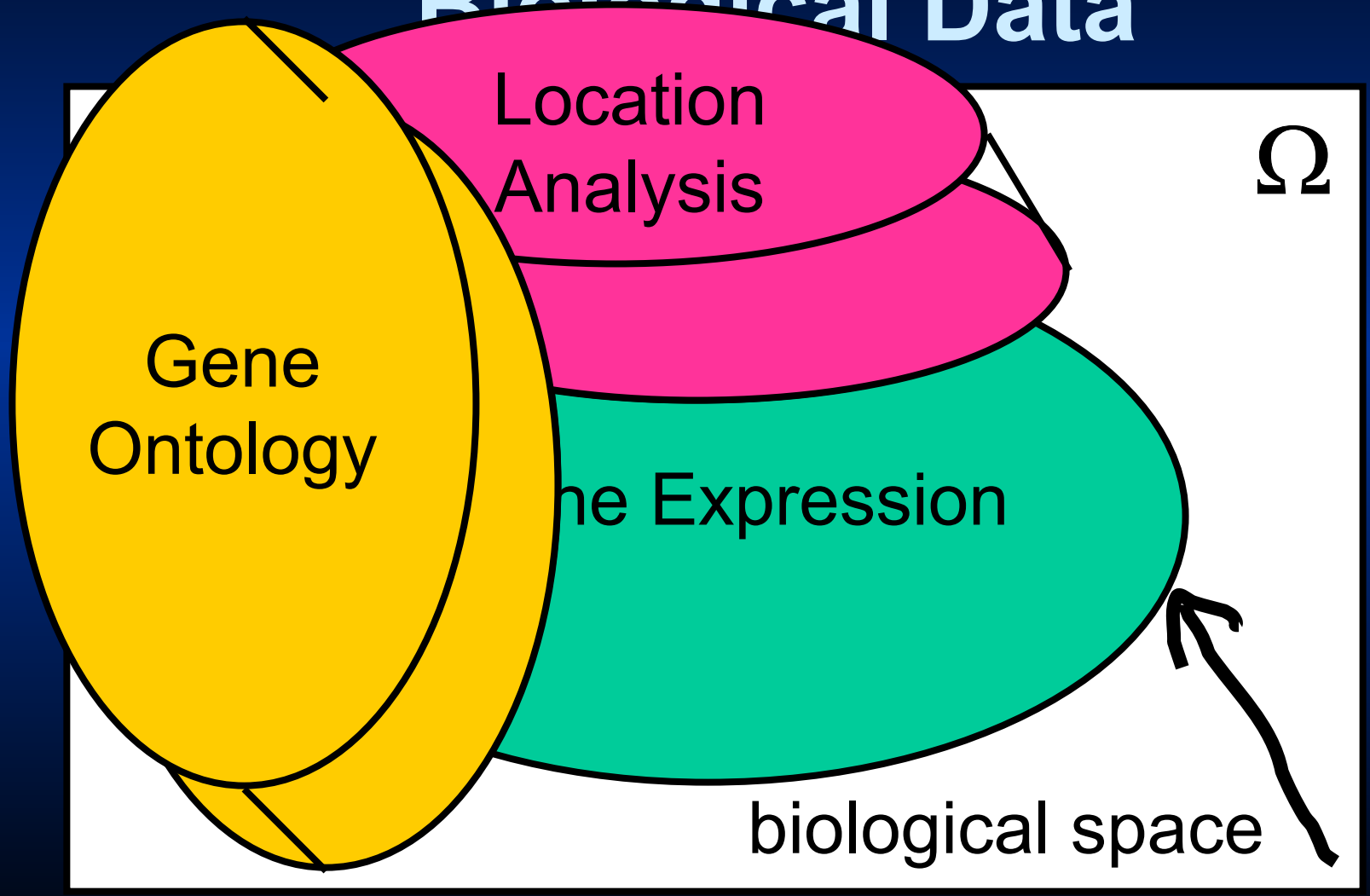


Computational Biology Department
Max-Planck-Institute for Molecular Genetics, Berlin

Motivation

Use additional large scale biological data to improve clustering of gene expression time-courses

Challenges of Heterogeneous Biological Data



Our approach

Semi-supervised

- Encode GO and location analysis as soft pairwise constraints
- Time-courses are modeled with a multivariate Gaussians
- Mixture estimation with constraints (Lange *et al.*, 2005, Lu and Leen, 2005)

Mixture Estimation with Constraints (1)

Maximize the complete likelihood:

$$P[X, Y|W, \Theta] = P[X|Y, \Theta]\, P[Y|W, \Theta]$$

where $\boldsymbol{X}$ is the observable data, $\boldsymbol{Y}$ the hidden data, Θ the model parameters, $W = \{W^+, W^-\}$ the pairwise constraints

The prior can be decomposed at:

$$P[Y|\Theta, W] = P[Y|\Theta]\, P[W^+|Y, \Theta]\, P[W^-|Y, \Theta]$$

Mixture Estimation with Constraints (2)

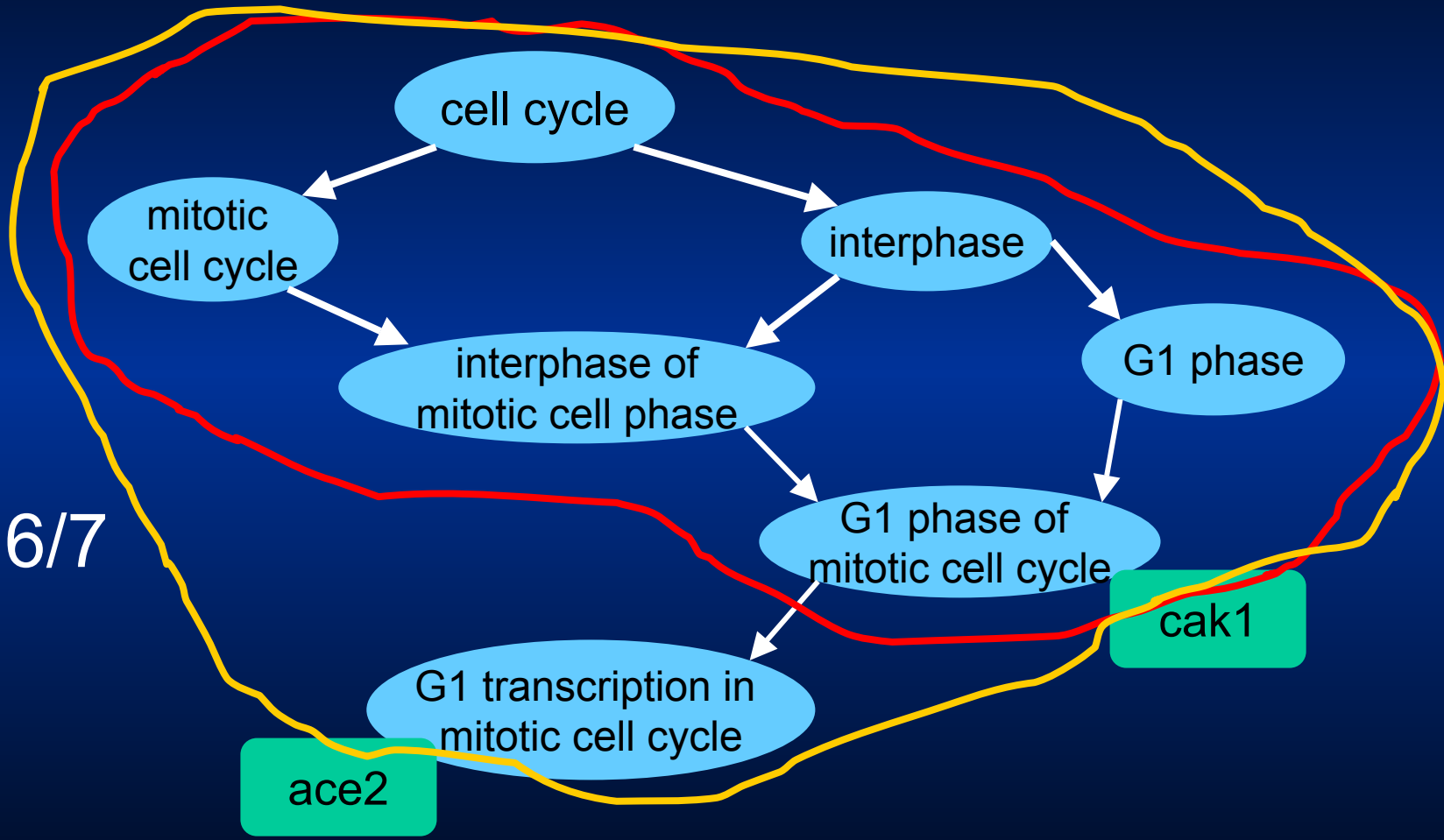
$$\mathrm{P}[W^{+} \mid Y, \Theta] \approx \exp \sum_{i} \sum_{j \neq i} -w_{ij}^{+} 1\{y_i \neq y_j\} \lambda^{+}$$

The posterior assignments are approximated by means of Gibbs sampling (Lu and Leen, 2005)

Data

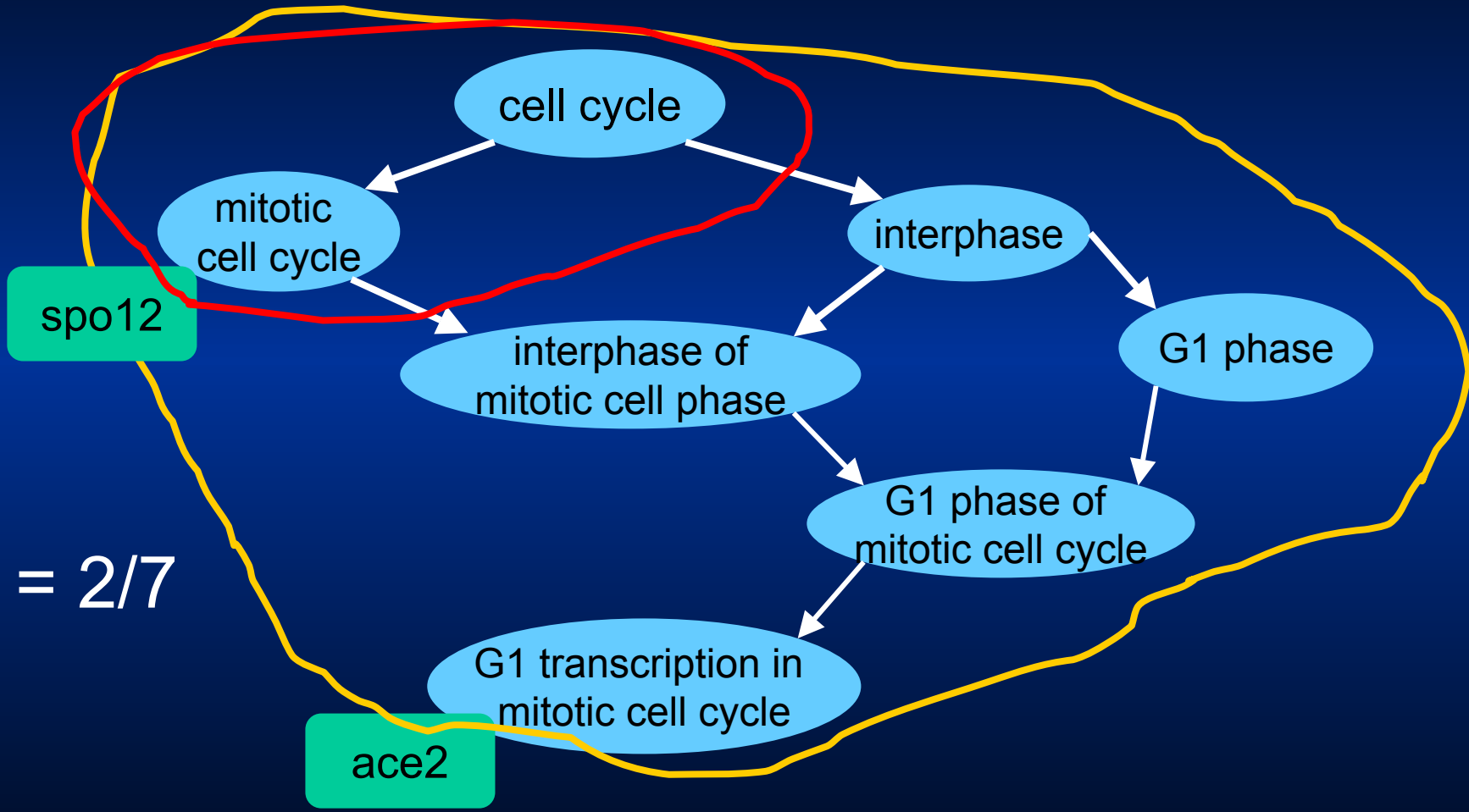
- Gene expression data
  - time-courses of 384 genes during mitotic cell division in Yeast (Cho, 1998)
  - expert classification into five cell-cycle phases
- Constraints
  - Gene Ontology (GO)
  - transcription factor location analysis (Lee, 2002)
  - true labels

Pairwise Constraints for GO



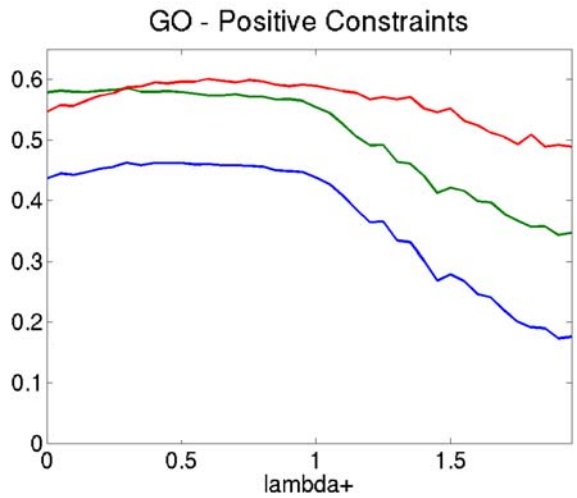
$w^+ = 6/7$

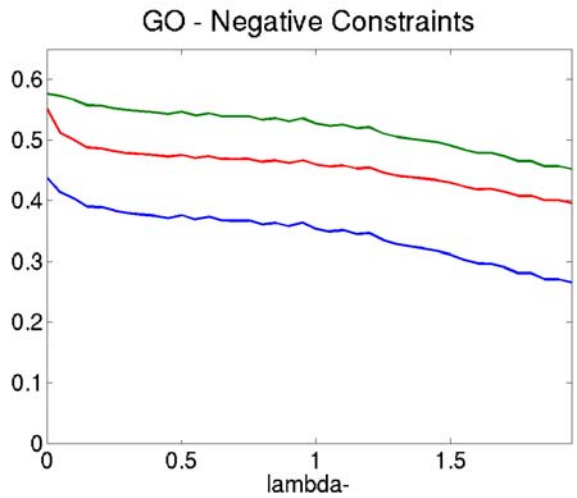
Pairwise Constraints for GO

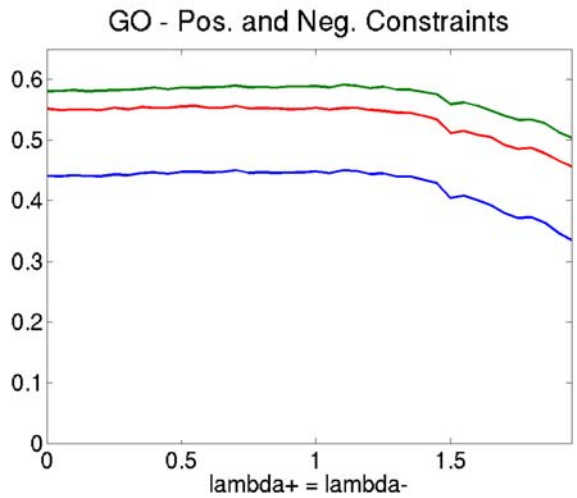


$w^+ = 2/7$

Constraints from GO



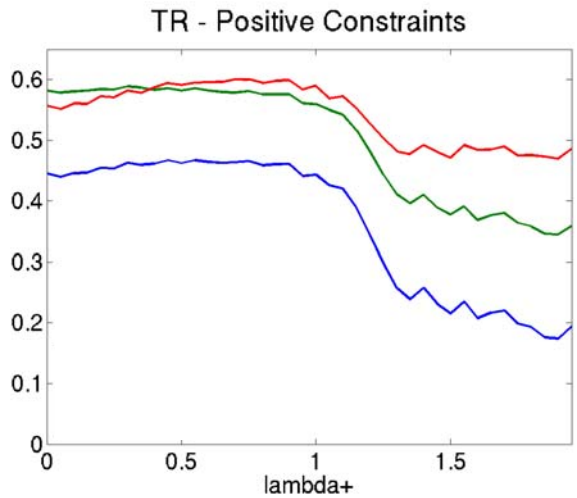


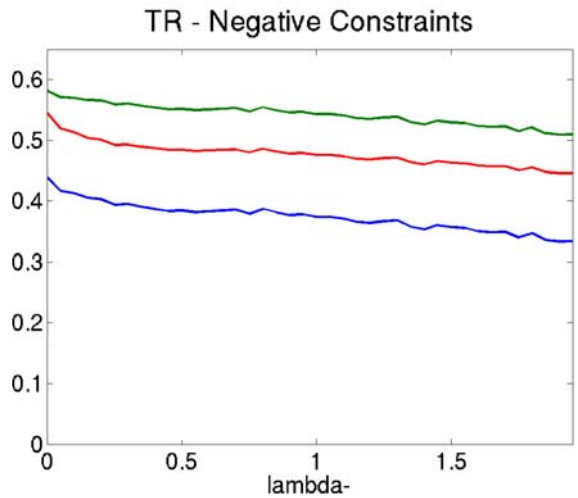


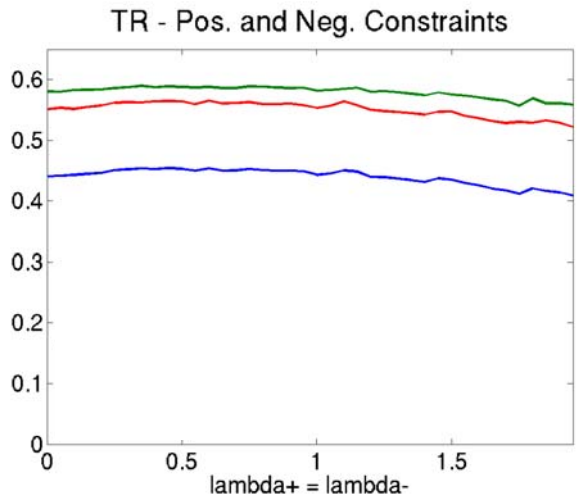
corrected Rand
specificity
sensitivity

51% of gene pairs constrained

Constraints from Location Analysis







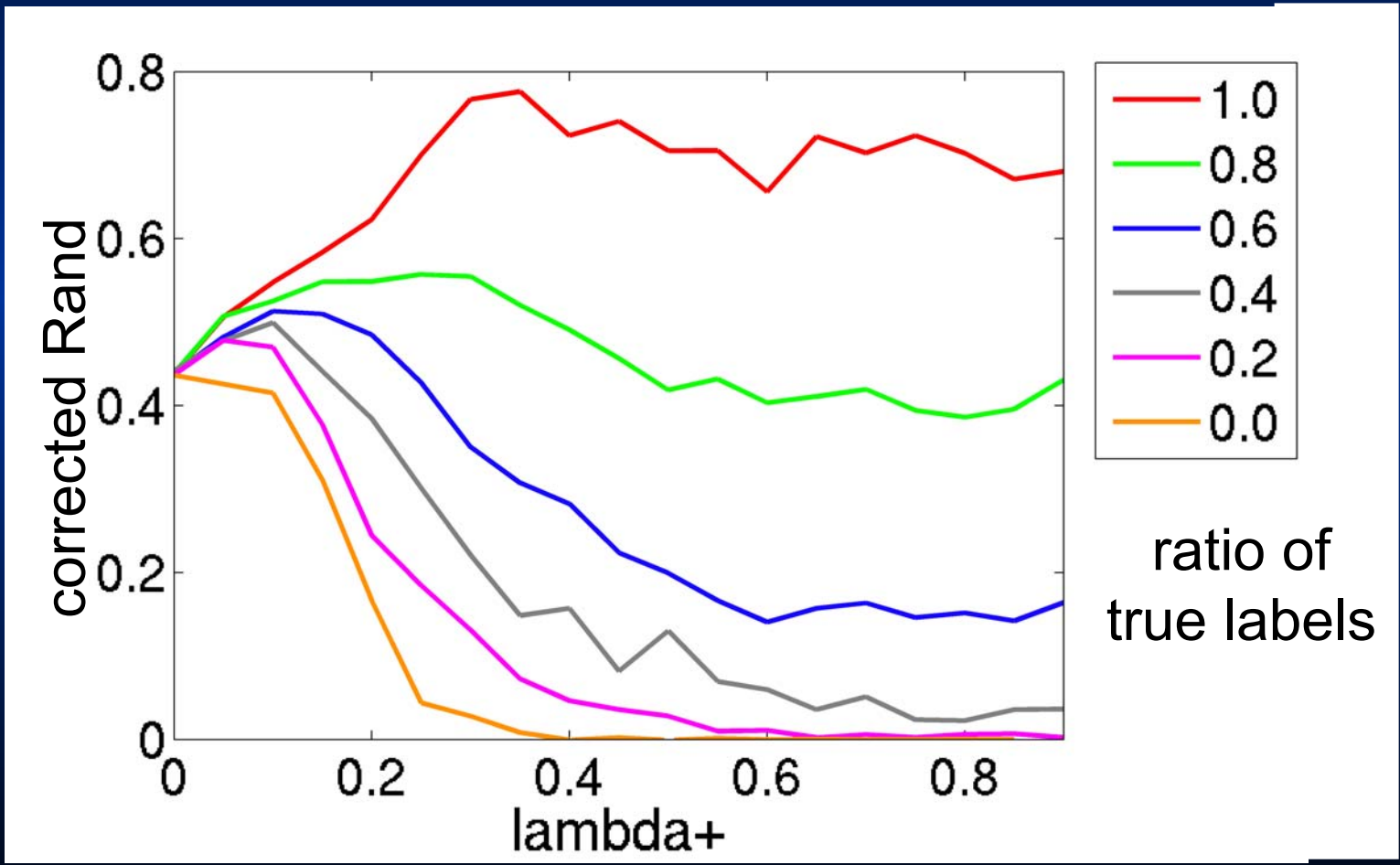
corrected Rand
specificity
sensitivity

40% of gene pairs constrained

Possible Explanations

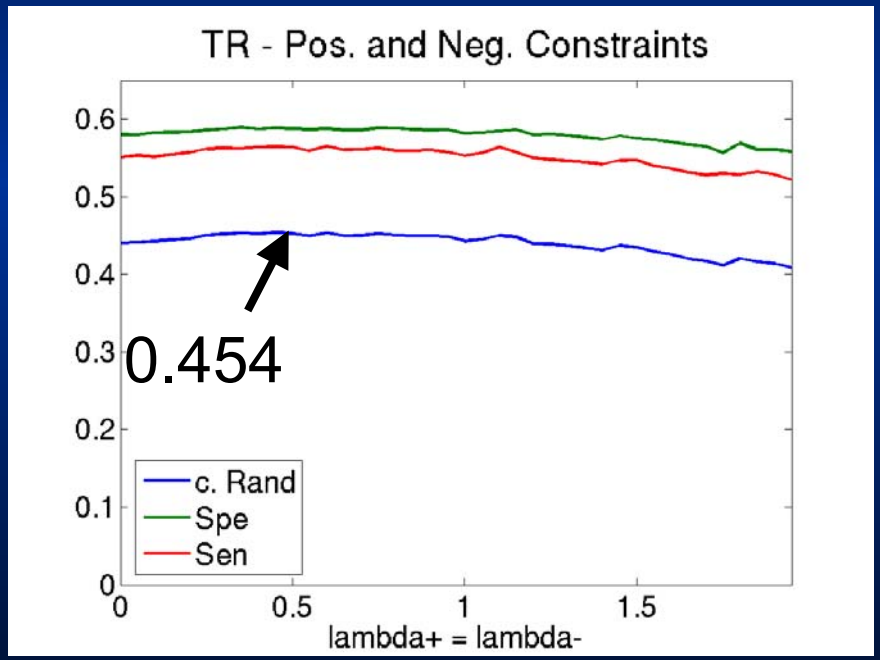
- Non-specific information content
- Noise in the data
- …

Constraints from True and Random Labels

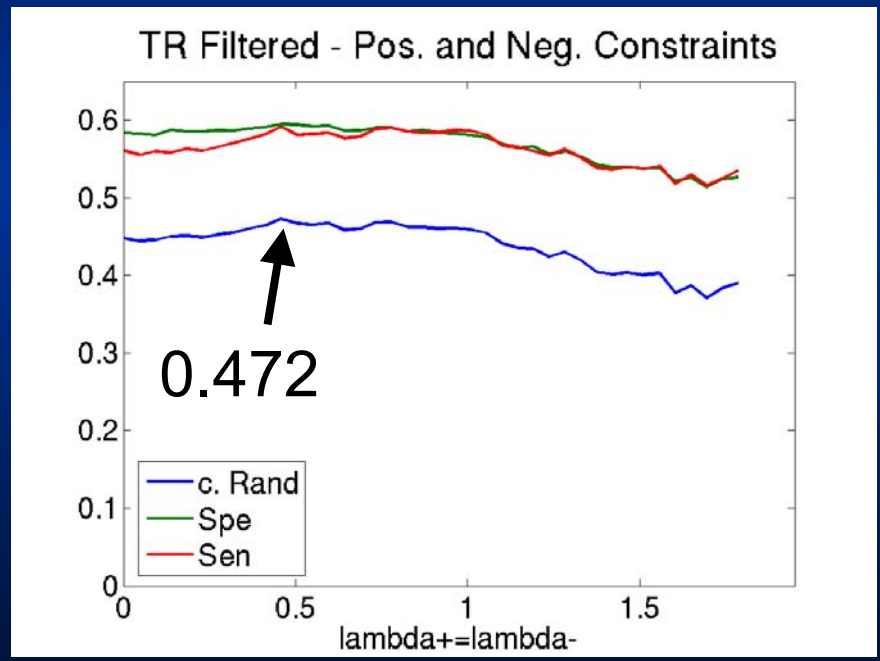


Filtered Constraints from Location Data

Non-filtered



Filtered



Summary and Outlook

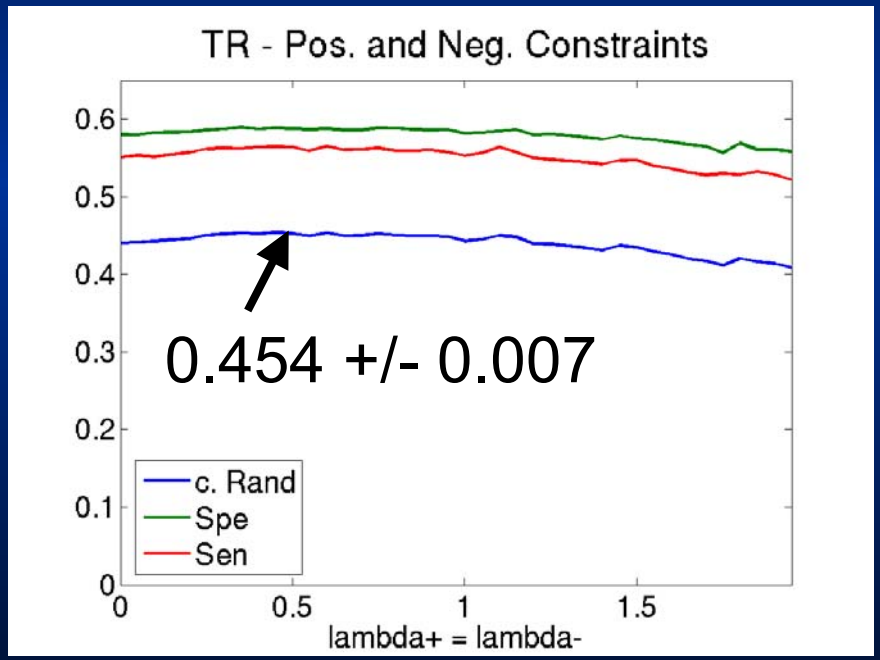
- A naïve analysis of heterogeneous data brings little improvements
- Data semantics?
  - GO does not yield further improvement
- Discovering relevant information?
  - learning of relevant constraints
  - find a right balance between λ+ and λ-

Thanks.

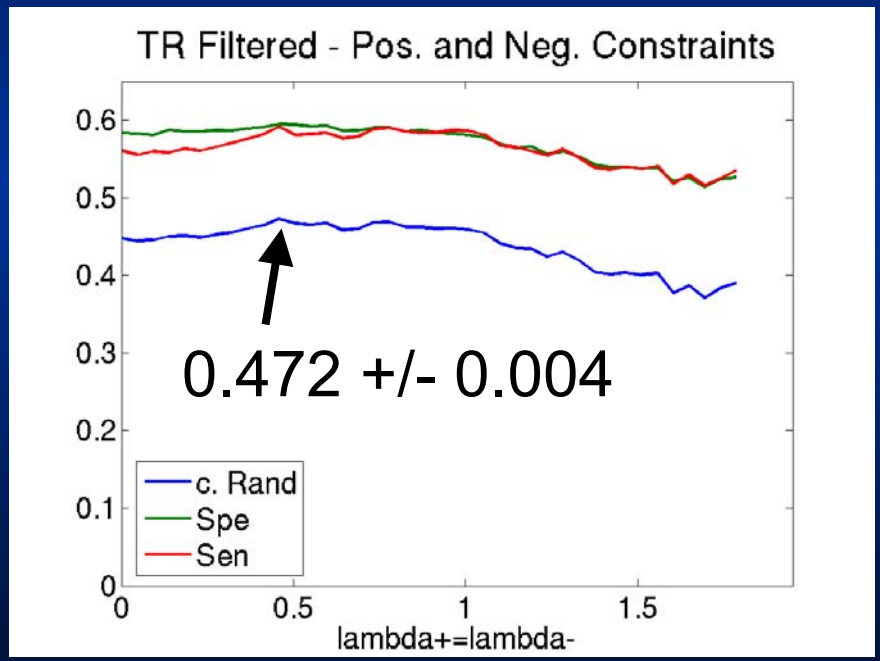
http://algorithmics.molgen.mpg.de

Filtered Constraints from Location Data

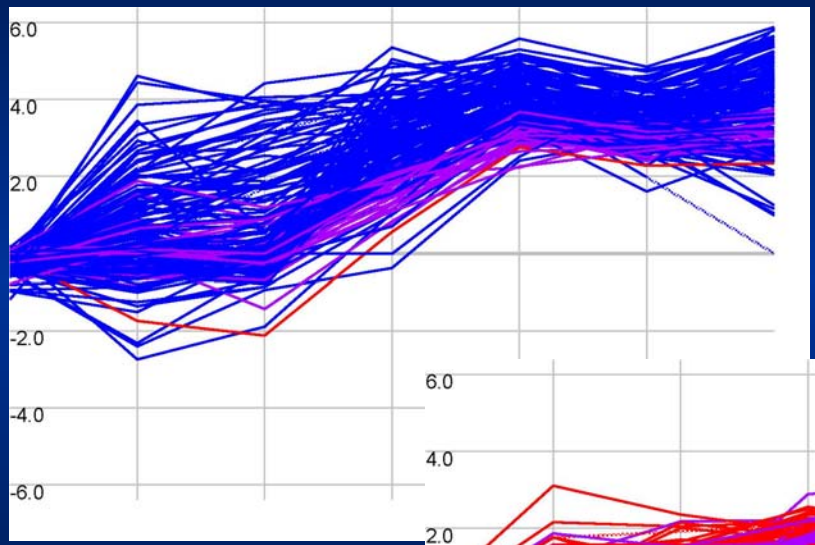
Non-filtered

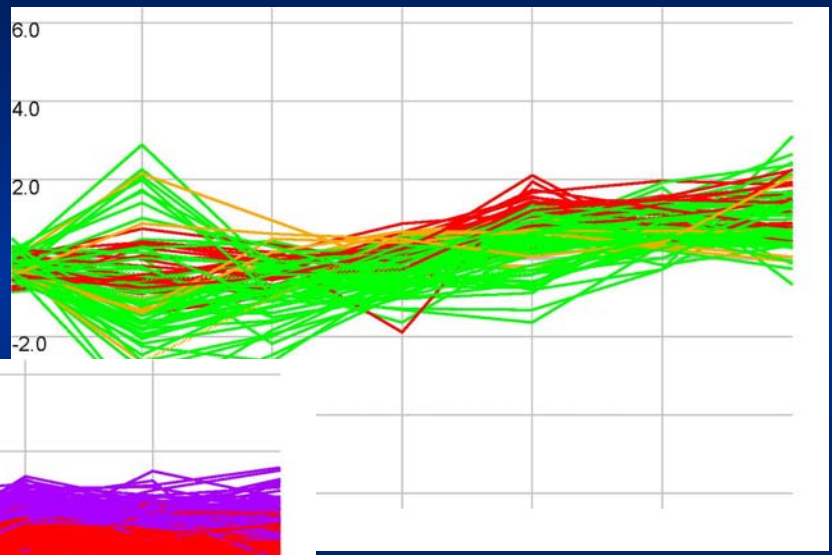


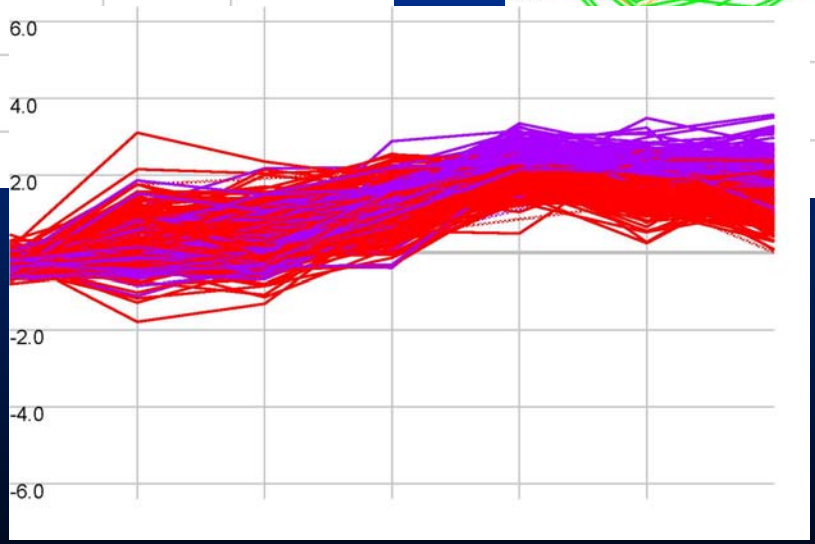
Filtered

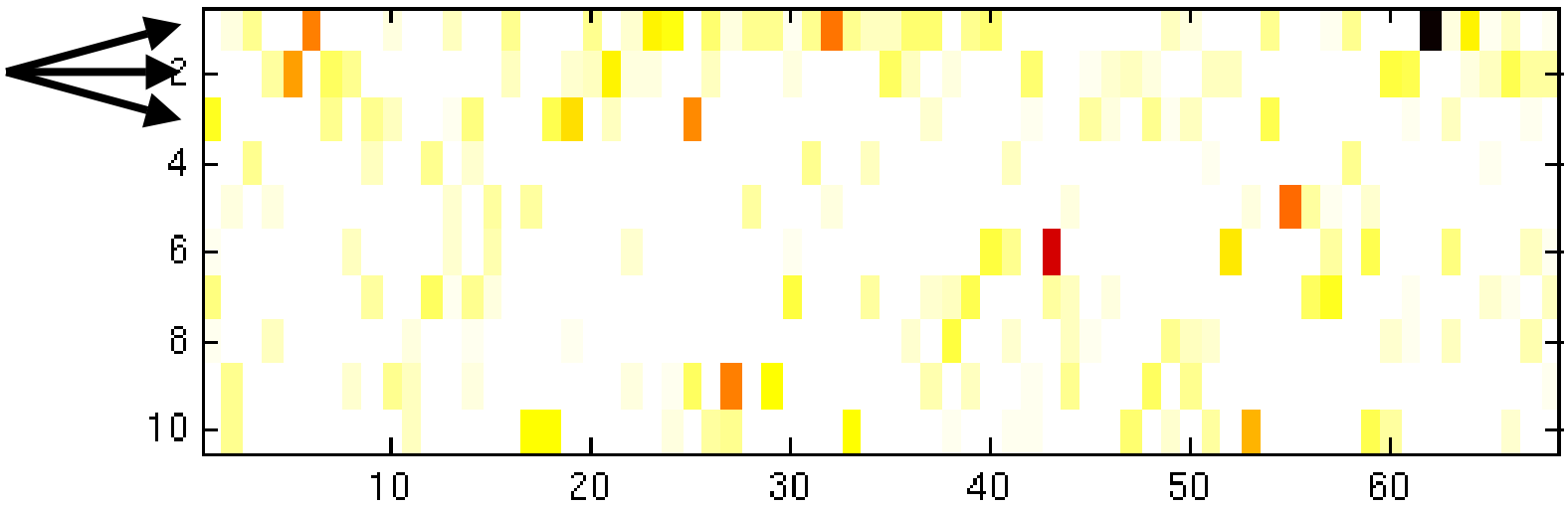


Examples - Sporulation data

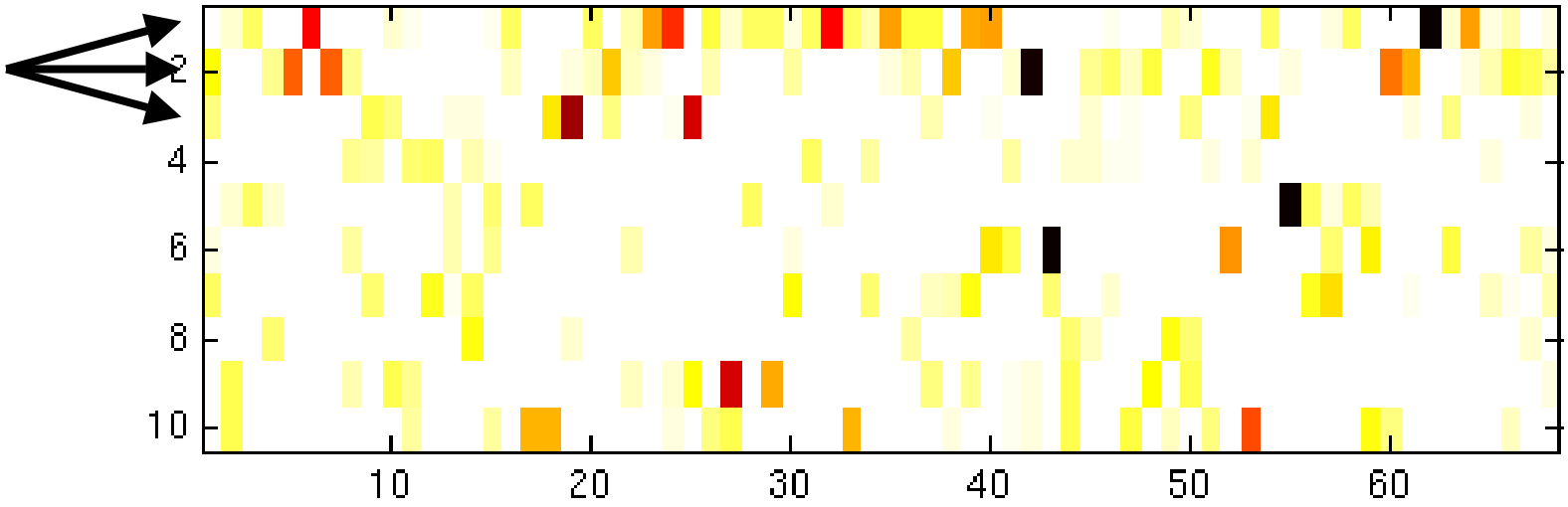








mixture estimation



mixt. est. with constraints

Pairwise Constraints for GO

We define the constraints for all pair of annotated genes:

$$w_{ij}^{+} = \frac{\#\{t_m \in Dag(g_i) \cap Dag(g_j)\}}{\#\{t_m \in Dag(g_i) \cup Dag(g_j)\}}$$

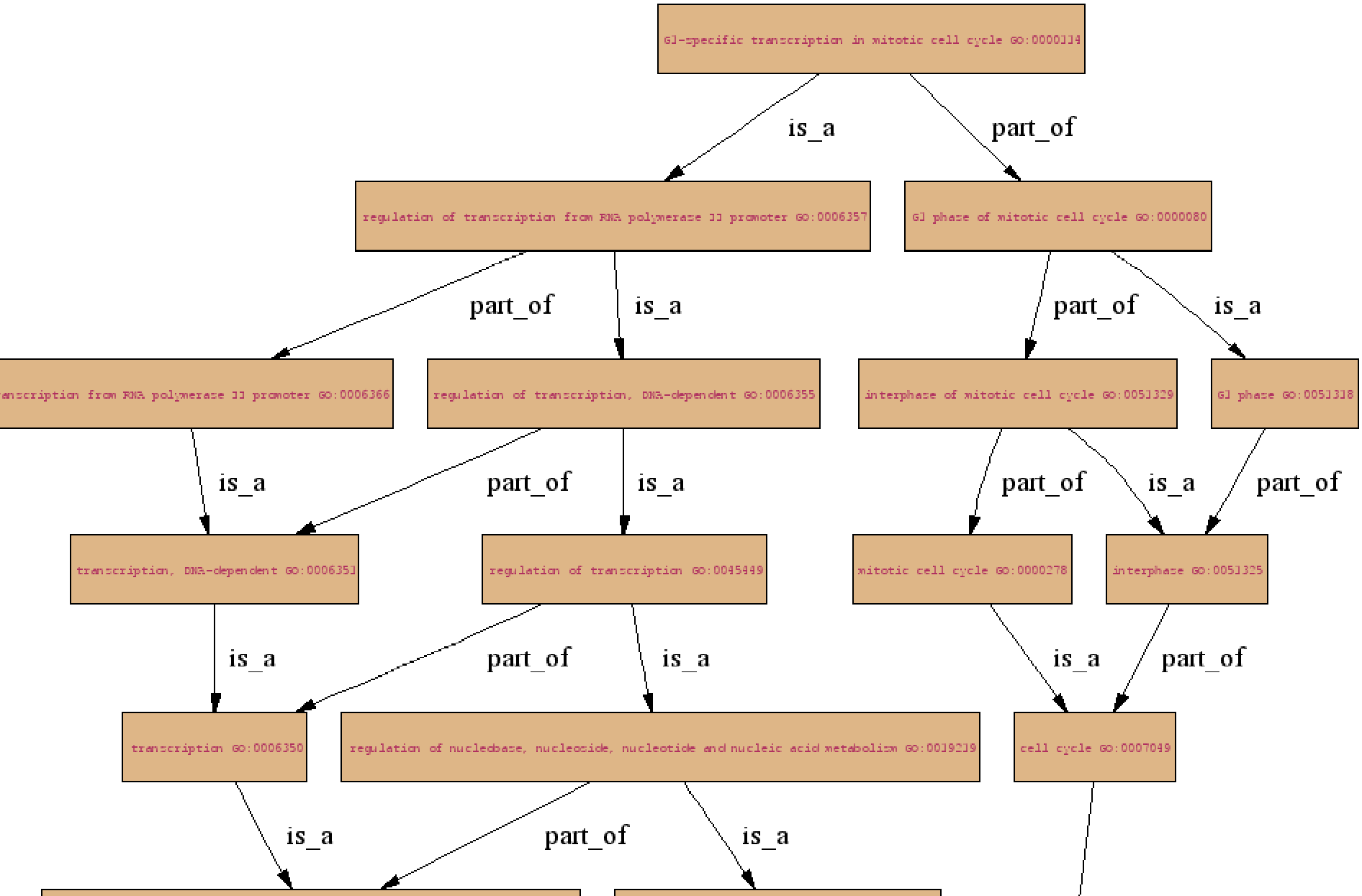
$$w_{ij}^{-} = \frac{\#\{t_m \in Dag(g_i) \cup Dag(g_j)\} - \#\{t_m \in Dag(g_i) \cap Dag(g_j)\}}{\#\{t_m \in Dag(g_i) \cup Dag(g_j)\}}$$

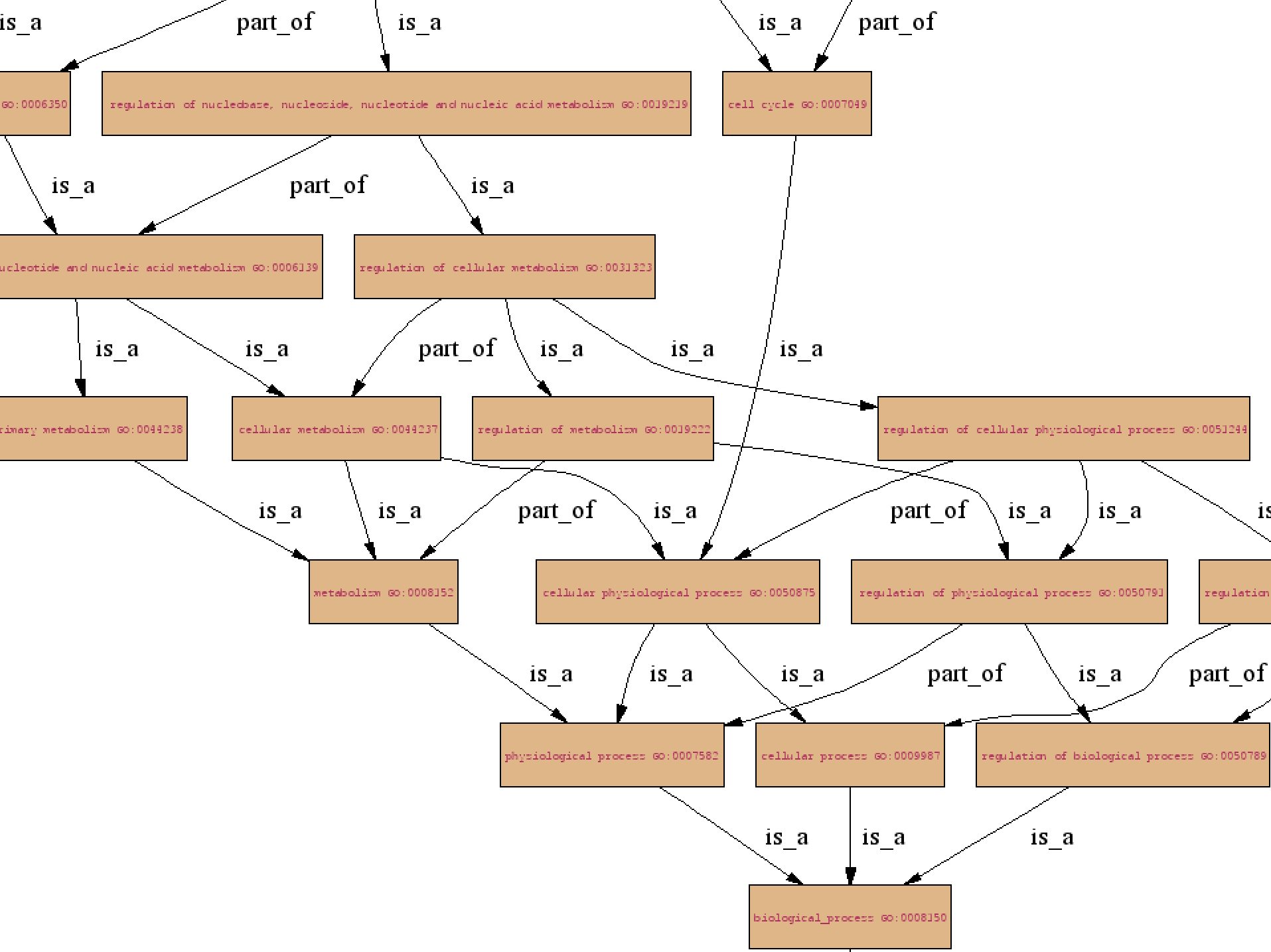
t_m are the terms in GO

$Dag(g_i)$ is the set of terms annotating g_i

More is less.

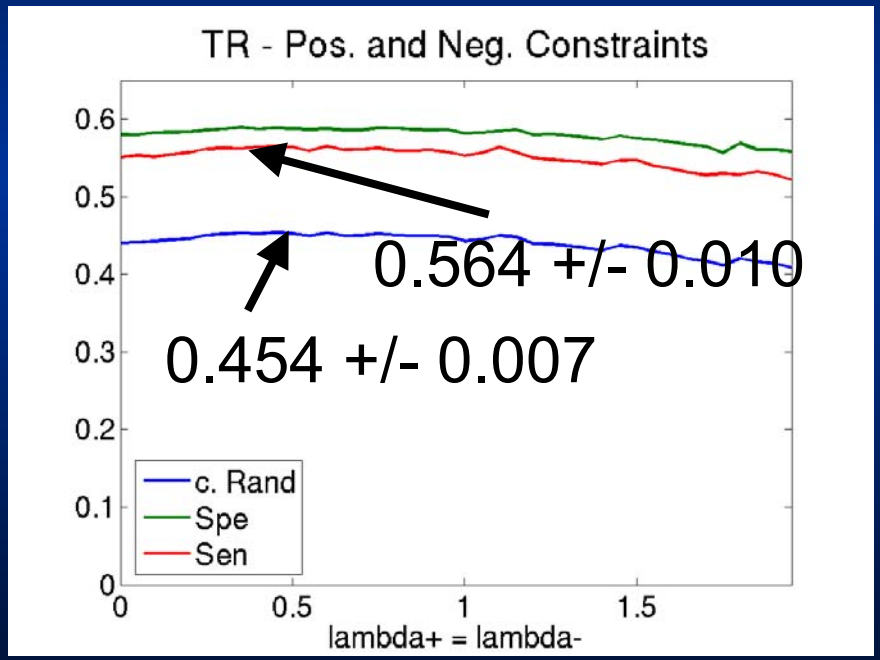
Sometimes …


G1-specific transcription in mitotic cell cycle GO:0000119
is_a
part_of
regulation of transcription from RNA polymerase II promoter GO:0006357
G1 phase of mitotic cell cycle GO:0000080
part_of
is_a
part_of
is_a
transcription from RNA polymerase II promoter GO:0006366
regulation of transcription, DNA-dependent GO:0006355
interphase of mitotic cell cycle GO:0051329
G1 phase GO:0051318
is_a
part_of
is_a
part_of
is_a
part_of
transcription, DNA-dependent GO:0006351
regulation of transcription GO:0045449
mitotic cell cycle GO:0000278
interphase GO:0051325
is_a
part_of
is_a
is_a
part_of
transcription GO:0006350
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism GO:0019219
cell cycle GO:0007049
is_a
part_of
is_a

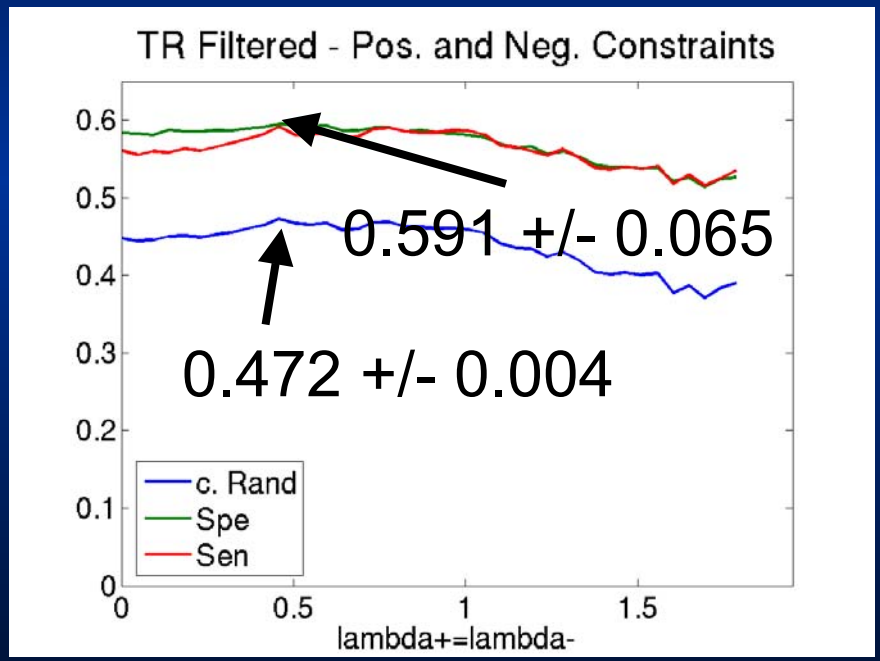

is_a
part_of
is_a
is_a
part_of
GO:0006350
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism GO:0019219
cell cycle GO:0007049
is_a
part_of
is_a
ucleotide and nucleic acid metabolism GO:0006139
regulation of cellular metabolism GO:0031323
is_a
is_a
part_of
is_a
is_a
is_a
rimary metabolism GO:0044238
cellular metabolism GO:0044237
regulation of metabolism GO:0019222
regulation of cellular physiological process GO:0051244
is_a
is_a
part_of
is_a
part_of
is_a
is_a
metabolism GO:0008152
cellular physiological process GO:0050875
regulation of physiological process GO:0050791
is_a
is_a
is_a
part_of
is_a
part_of
physiological process GO:0007582
cellular process GO:0009987
regulation of biological process GO:0050789
is_a
is_a
is_a
biological_process GO:0008150

Expert Filter of Constraints Transcription Factors

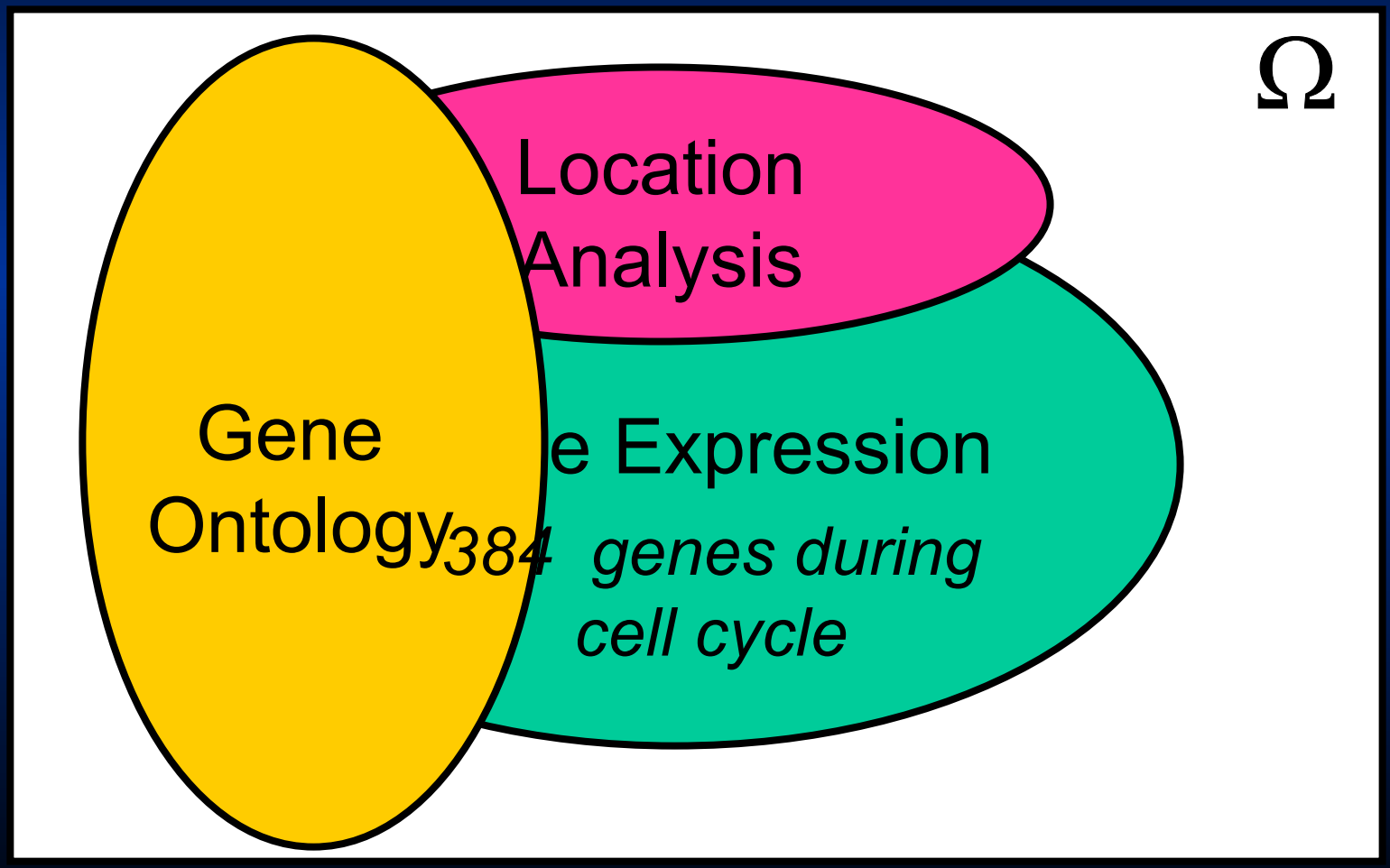
Non-filtered



Filtered



Data


Ω
Location Analysis
Gene Ontology
e Expression
384 genes during cell cycle

Mixture Estimation with Constraints (3)

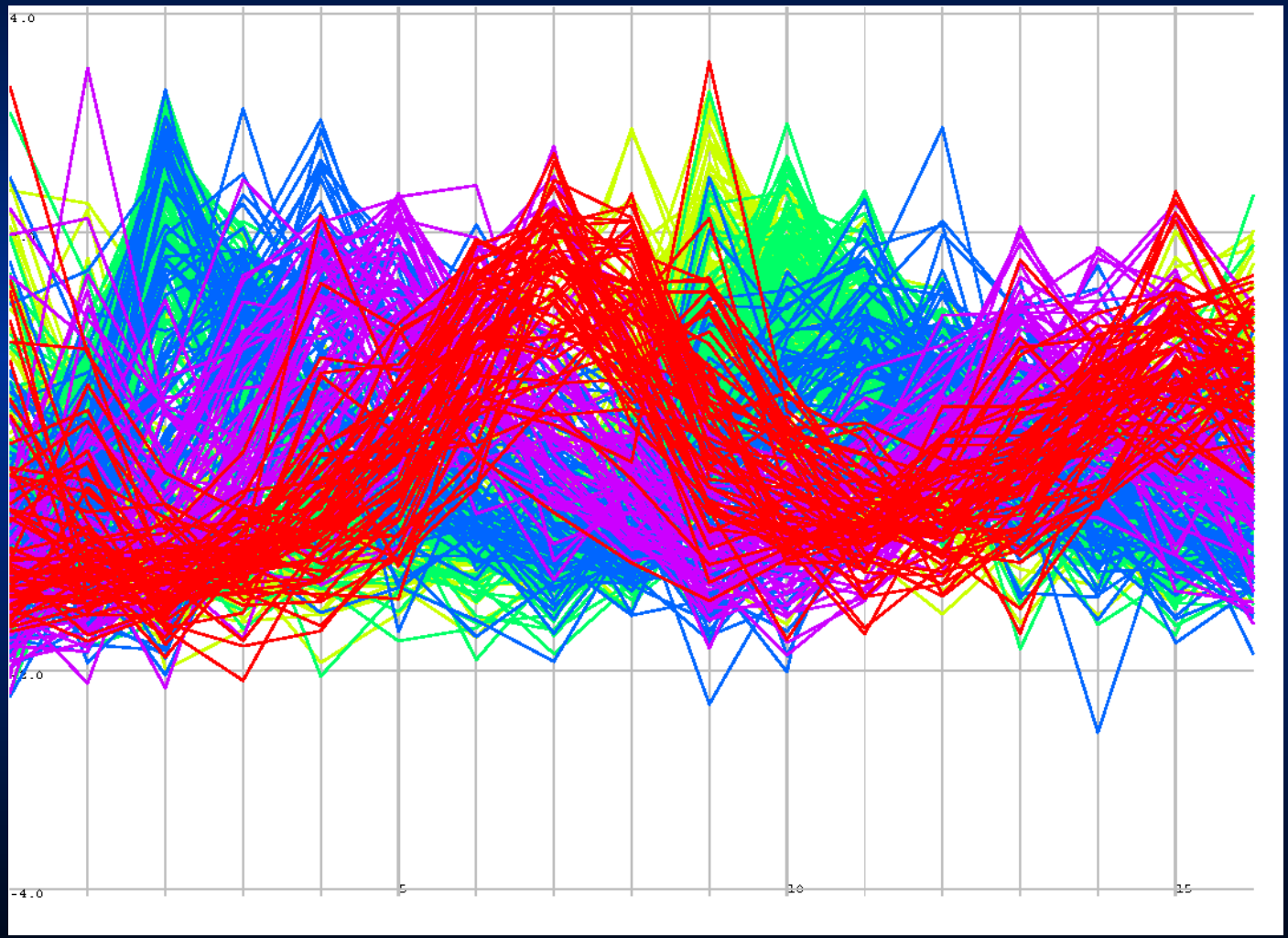
The posterior P[y_i=k|Y_i',X,W,θ_k] is approximated by means of Gibbs sampling :

$$\alpha_k \mathrm{P}[x_i \mid \theta_k] \exp \sum_{j \neq i} -\lambda^+ w_{ij}^+ (1 - \mathrm{P}[y_j = k \mid Y_i', X, W, \theta_k])$$

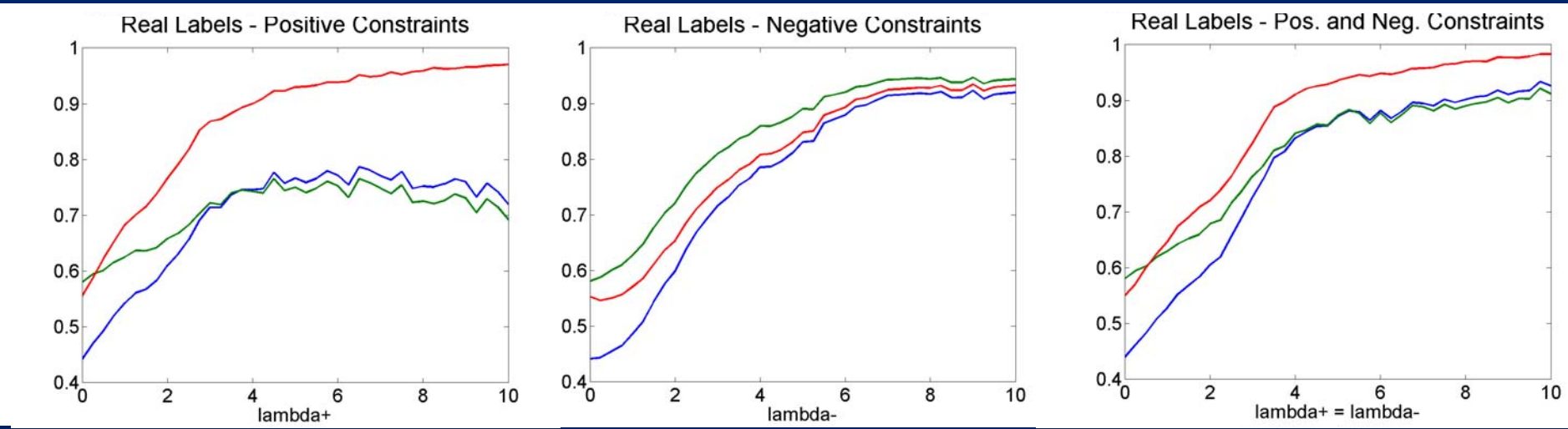
where

$$Y_i' = \{y_1^t, \ldots, y_{i-1}^t, y_{i+1}^{t-1}, \ldots, y_N^{t-1}\}$$

Yeast Cell Cycle



Constraints from True Labels



corrected Rand
specificity
sensitivity

5% of gene pairs constrained