b-it Bonn-Aachen International Center for Information Technology

universitätbonn

Boosting Statistical Network Inference by Incorporating Prior Knowledge from Multiple Sources

Paurush Praveen

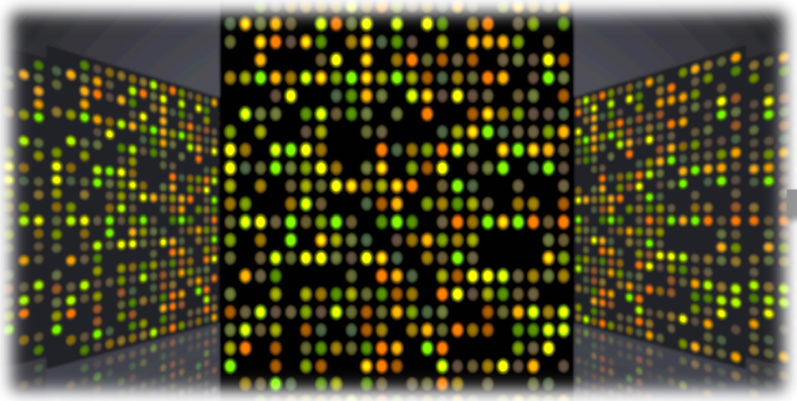
Department of Algorithmic Bioinformatics, Bonn-Aachen International Center for IT (B-IT), Dahlmannstr. 2, 53113 Bonn

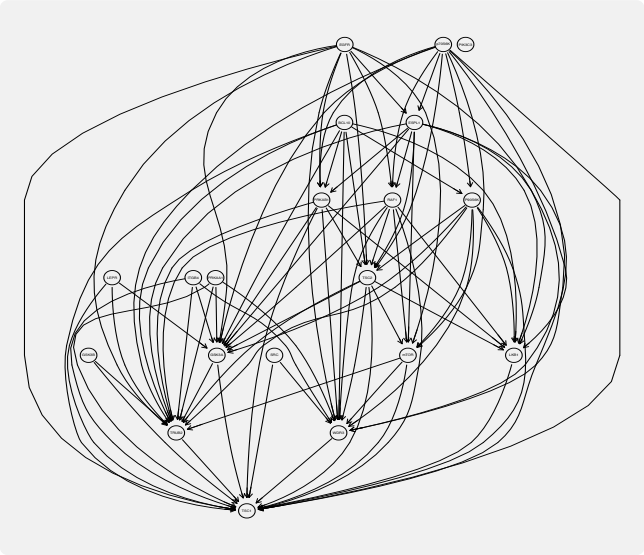
Overview

Boosting Statistical Network Inference by Incorporating Prior Knowledge from Multiple Sources

- Motivation
- Approach (LFM and NOM)
- Test and validation
- Application

Motivation

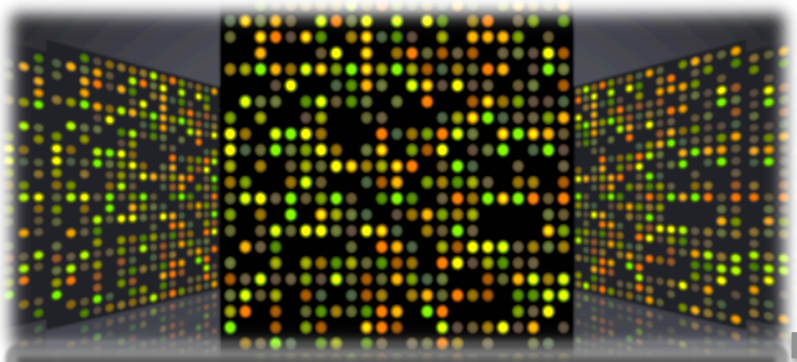




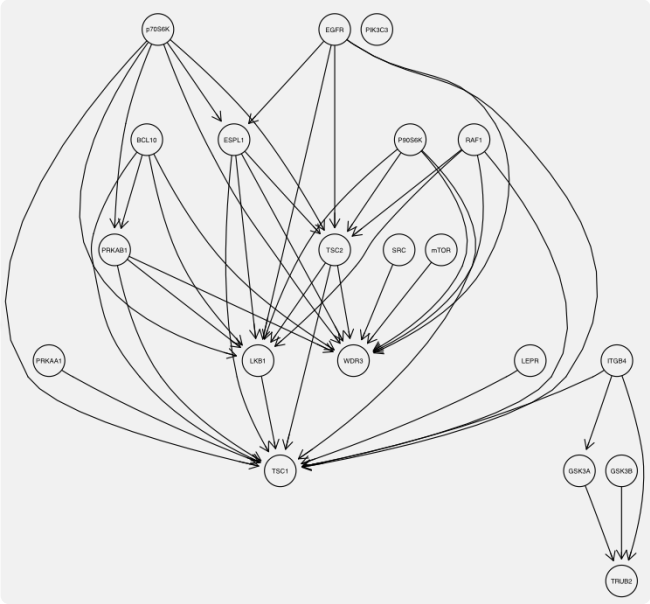
Inherent noise data
Low sample size

High false positives and false negatives.

Motivation



Knowledge from databases, ontologies, literature etc.



A quantified prior from existing biological knowledge and information together with data

- ✓ Reduced false positive and false negative
- ✓ Network closer to real world

Problem Definition

- Heterogeneous Biological Information
  - *Interaction data*
  - *Ontology information*
  - *Literature information*
- Missing information
  - *Missing annotations*
  - *Incomplete interaction data*
- Numerical handling of knowledge

Prior

- Bayes Rule

$$p(\phi \mid D) = \frac{p(D \mid \phi)p(\phi)}{p(D)}$$

- *Example:* $P(\Phi) = \prod_{i,j} P(\Phi_{ij})$

$$P(\Phi_{ij} \mid \nu) = \frac{1}{2\nu} \exp\left(\frac{-|\Phi_{ij} - \hat{\Phi}_{ij}|}{\nu} \right)$$

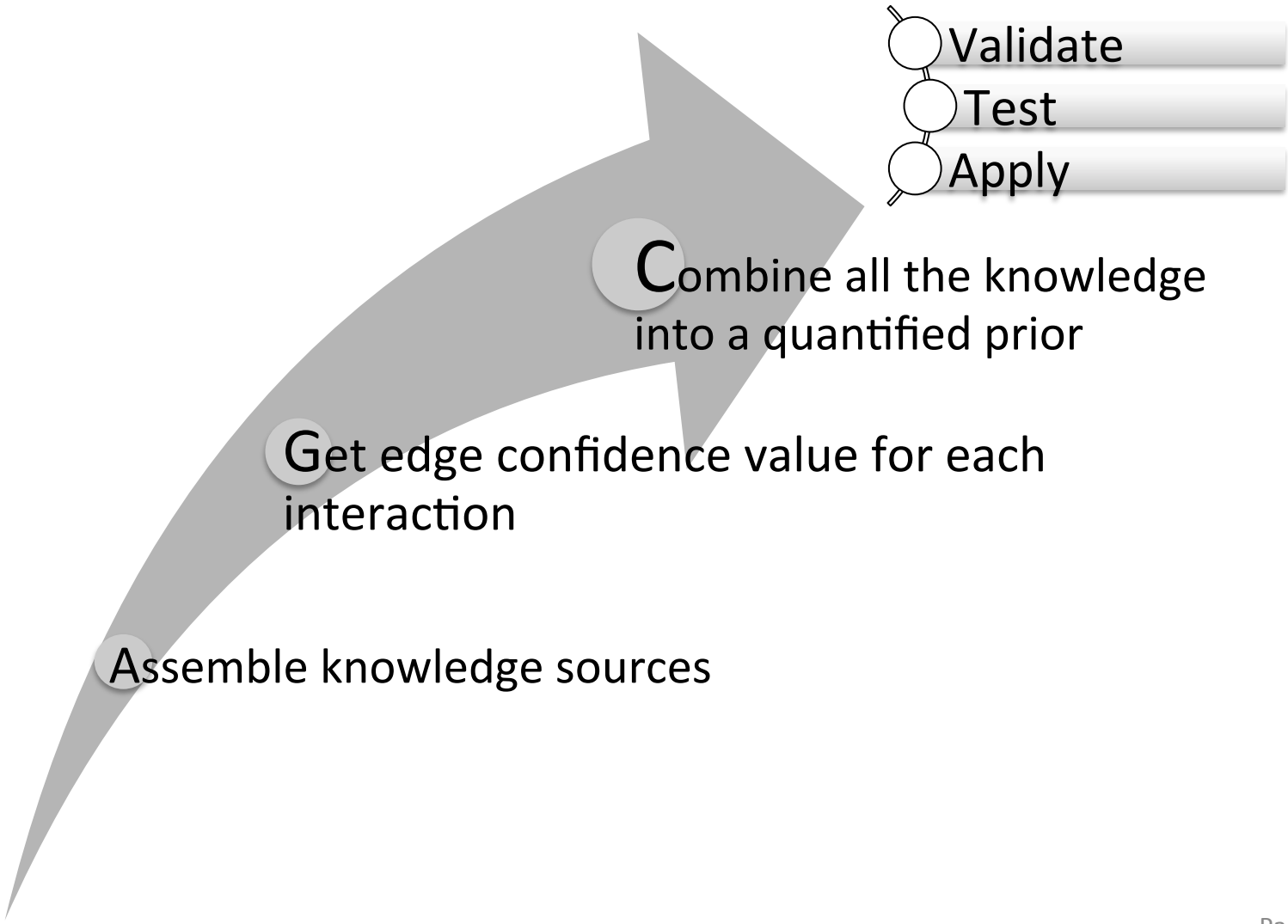
$\hat{\Phi}$ = Prior belief matrix

ν = Hyperparameter of Laplace Distribution

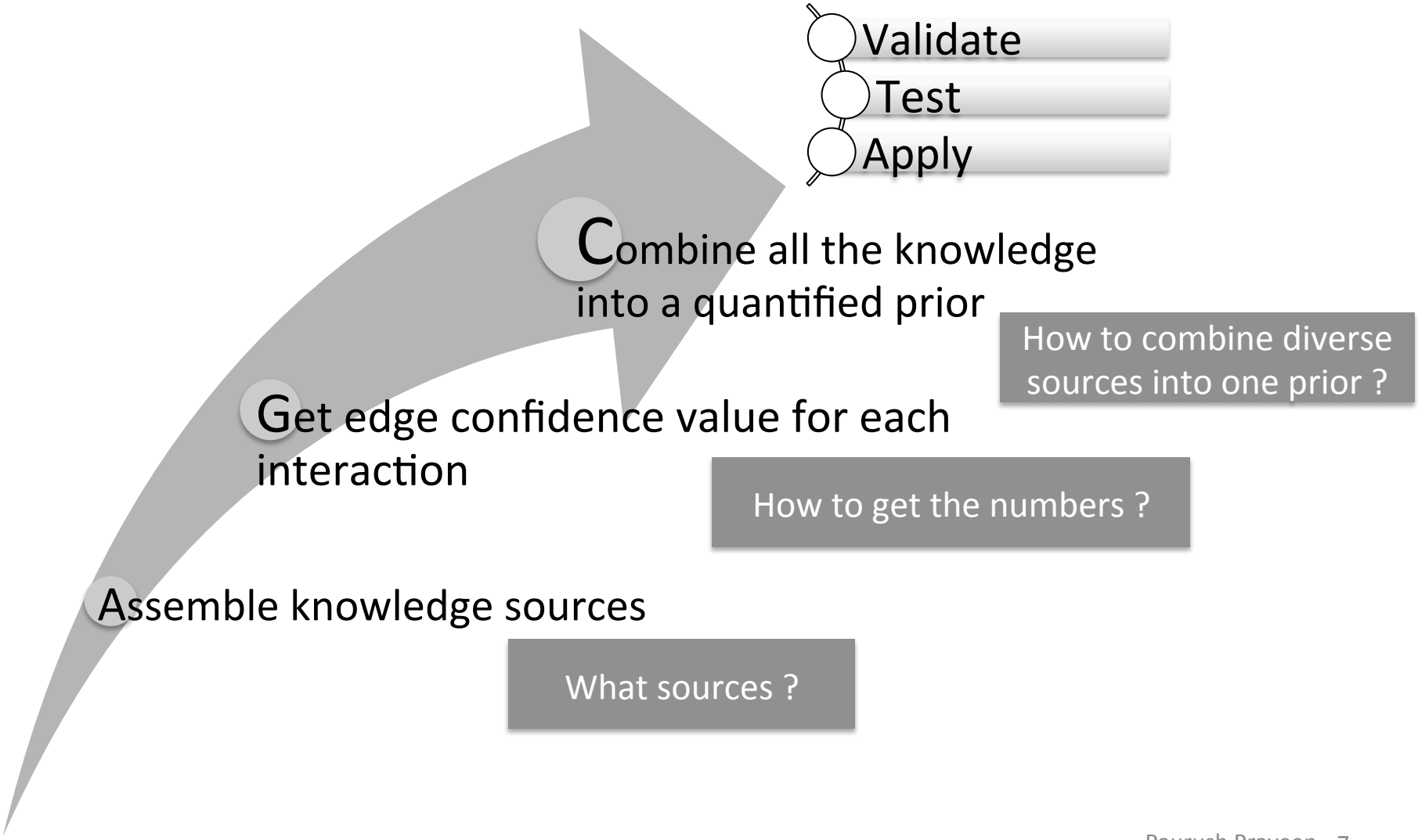
(Fröhlich et al., 2007, 2008)

- Biases the learning to consider known interactions.

Idea


Validate
Test
Apply
Combine all the knowledge into a quantified prior
Get edge confidence value for each interaction
Assemble knowledge sources

Idea


Validate
Test
Apply
Combine all the knowledge into a quantified prior
How to combine diverse sources into one prior ?
Get edge confidence value for each interaction
How to get the numbers ?
Assemble knowledge sources
What sources ?

Information Sources

- Interaction Databases
  - PathwayCommons, String, HPRD *etc.*
- Gene Ontologies
  - Consider similarity in annotation of two genes
- Protein Domain Information
  - Similarity in domains of interacting protein
- Interacting Domain Information
  - Domains that can interact in the proteins
- Literature Data
  - Interaction quoted in existing literature

Edge Confidence

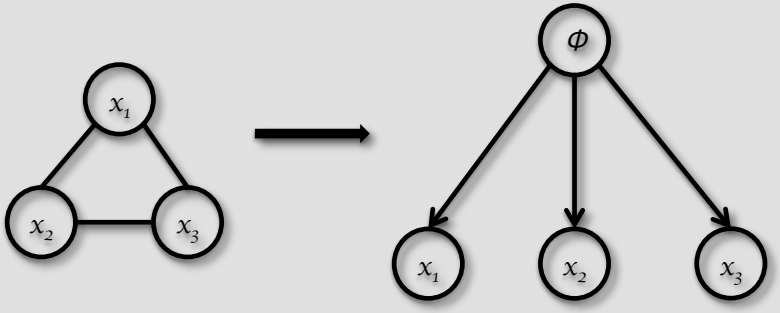
Confidence in the edge or interaction from different sources

- Interaction graph from databases or literature
  - 1/Shortest path length (To be updated to Graph Diffusion Kernel)
- Ontologies (GO terms)
  - Similarity in GO terms (Lin *et.al.* 1989)
- Protein domain data
  - Domain similarity (*cosine similarity of protein domain signature vector*)
- Interacting domain data
  - Occurrence of interacting domains pairs in the query protein

Latent Factor Model (LFM)

Assumption: Highly related knowledge sources originate from a true but unknown network

- Correlated interaction networks originate from the true network
- The true network is a latent variable Φ
- We intend to use the information from knowledge sources to learn Φ



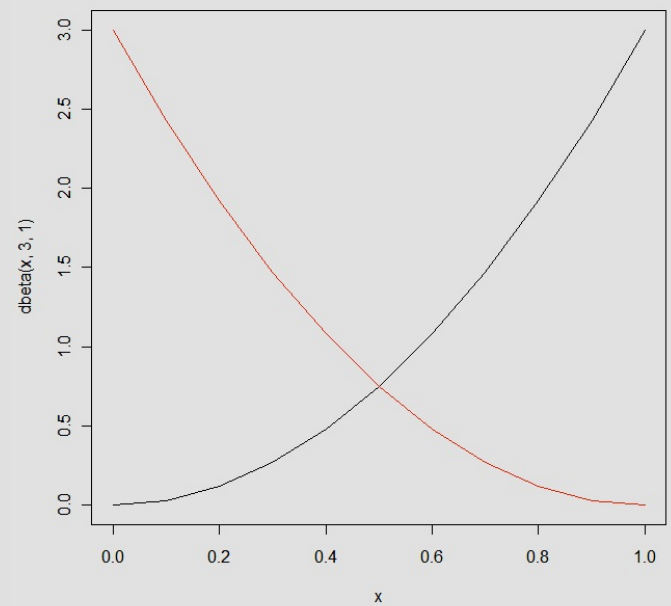
Where,
x_1, x_2, x_3 are correlated variables

Φ is the latent variable and parent for x_1, x_2, x_3

$$P(\phi|x_1, x_2 ... x_n) = \frac{\prod P(x_i|\phi)P(\phi)}{P(x_1, x_2, \ldots x_n)}$$

Latent Factor Model

- We assume that the confidence values from different sources follow a beta distribution (1)
- The log likelihood of the network can be given as (2)
- Learning by *"Nested" MCMC* sampling using *Metropolis Hastings Algorithm*
- *'Nested'* MCMC
  - Edge insertion, deletion, and reversal
  - Modification of α, β *every* i^{th} *iteration*



$$P(x_{i_{ab}}|\phi_{ab}=1) \sim Be(x_{i_{ab}},\alpha,1); \alpha \geq 2$$

$$P(x_{i_{ab}}|\phi_{ab}=0) \sim Be(x_{i_{ab}},1,\beta); \beta \geq 2$$

(1)

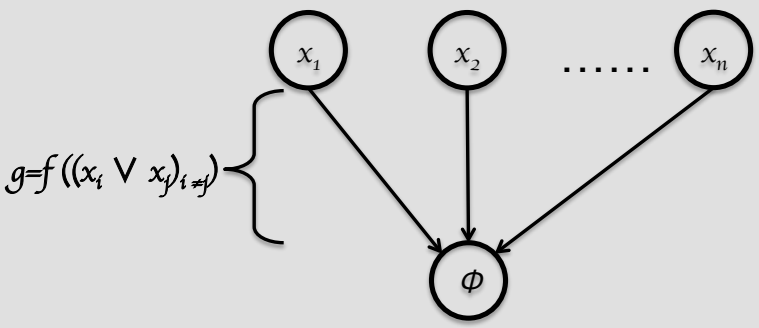
$$logP(x_1,x_2\ldots x_n|\phi_{ab}) = \sum_s \sum_{a,b} \begin{cases} log(Be(x_{i_{ab}},\alpha,1)), \phi_{ab}=1 \\ log(Be(x_{i_{ab}},1,\beta)), \phi_{ab}=0 \end{cases}$$

(2)

Noisy-OR Model (NOM)

Assumption: Given a set of causes x_i for an effect Φ the probability of each cause to be sufficient to prove the effect is independent of presence of other causes.

- Source of information x_i is the cause
- Overall prior knowledge is effect Φ.
- The cause x_i consists of confidence for every edge (P_{ab}) in the network



Where,

x_1, x_2 … x_n are sources of information

Φ is the effect and combination of all the information listed above using the function ***g***

The Noisy-OR function given as

$$P(\phi|x_i) = 1 - \prod_{x_i} 1 - P_{ab_{x_i}}$$

Noisy-OR Model

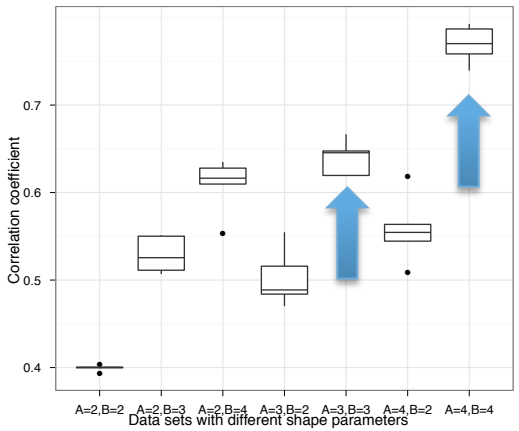
- We apply the Noisy-OR model on the sources of information to get the overall prior knowledge.

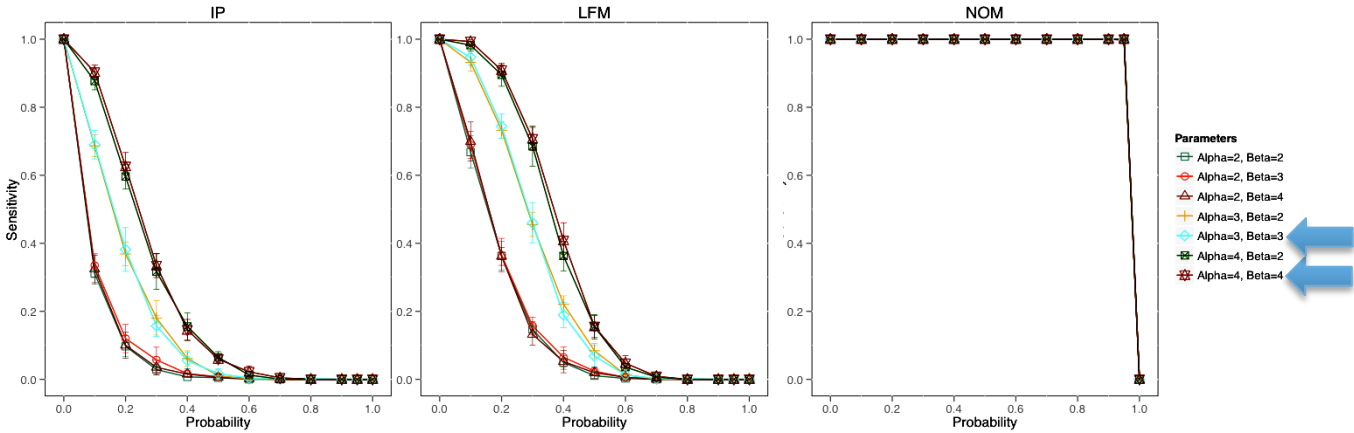
- The knowledge sources have the confidence for each possible edge in the desired set of genes

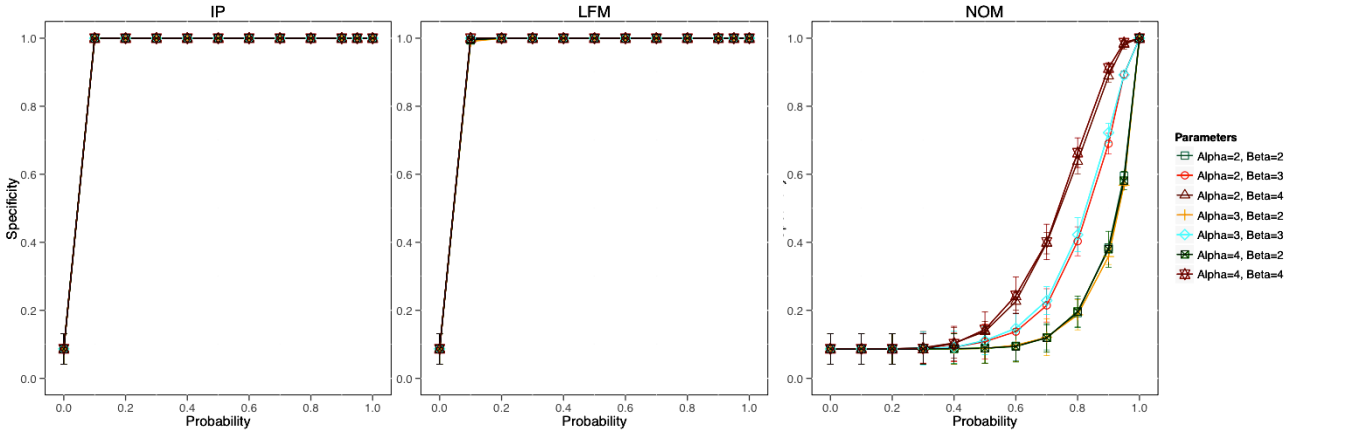
Both the methods result into a matrix with values ranging between 0 and 1

Simulations I

Value Range and Data Correlation



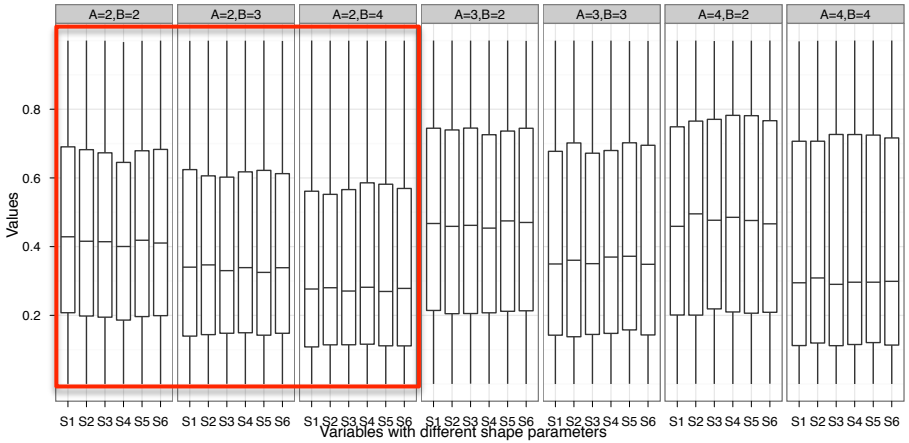


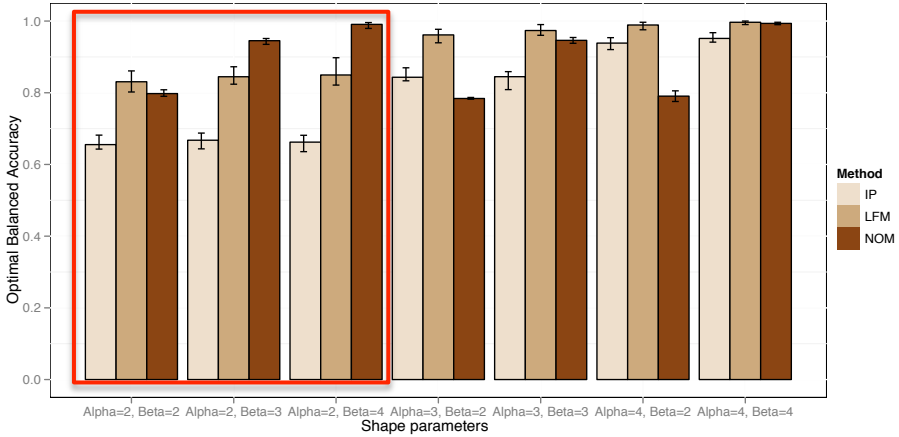


- With higher correlation among the information sources LFM and NOM perform better

Simulations I

Value Range and Data Correlation

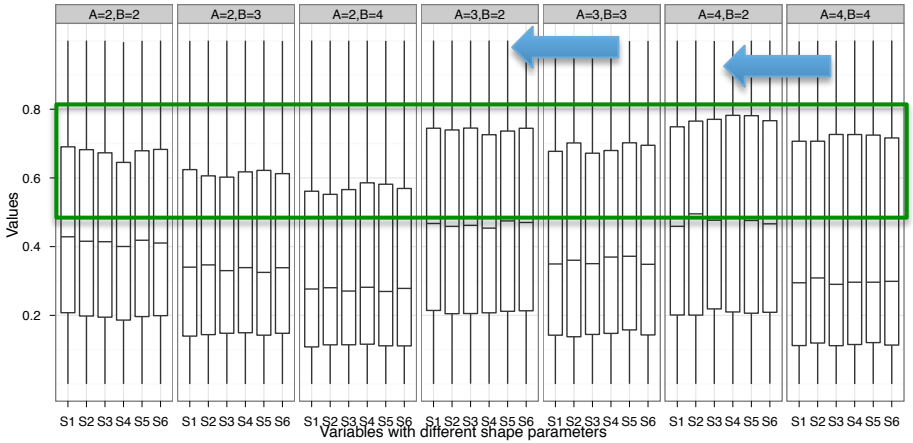


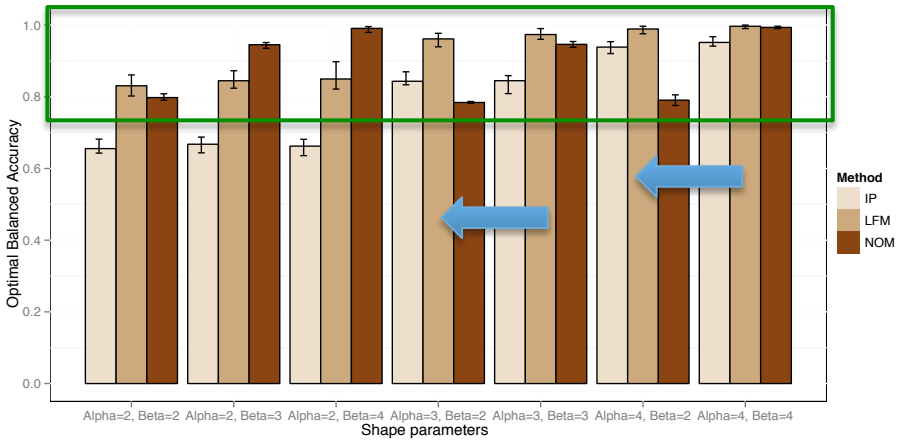


- The IP method shows poor performance if the information data has low content compared to LFM and NOM

Simulations I

Value Range and Data Correlation

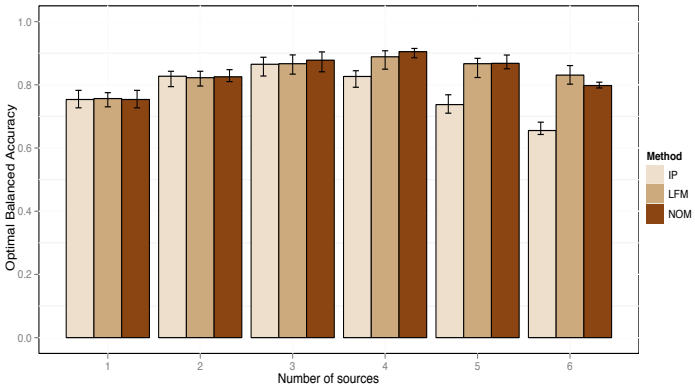


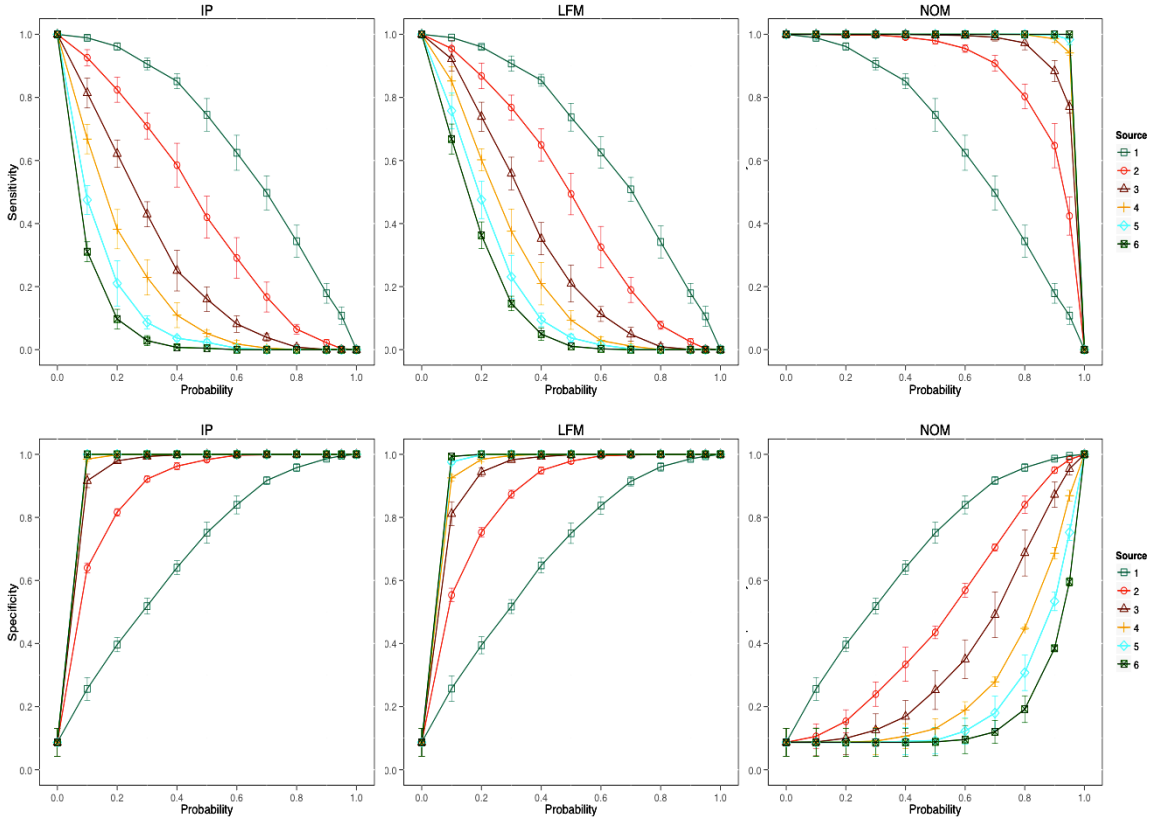


- The LFM and NOM methods maintain their performance even if the information content is not very strong
- NOM performs well in case of weak information
- LFM outperforms the IP method throughout this test

Simulations II

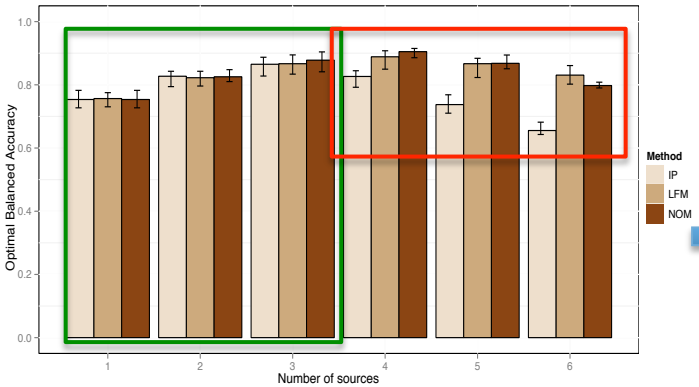
Number of Sources

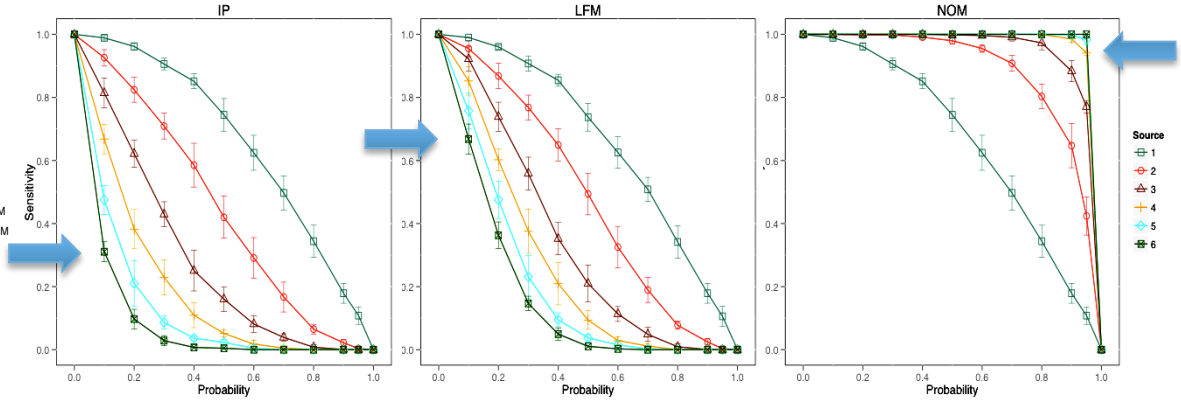


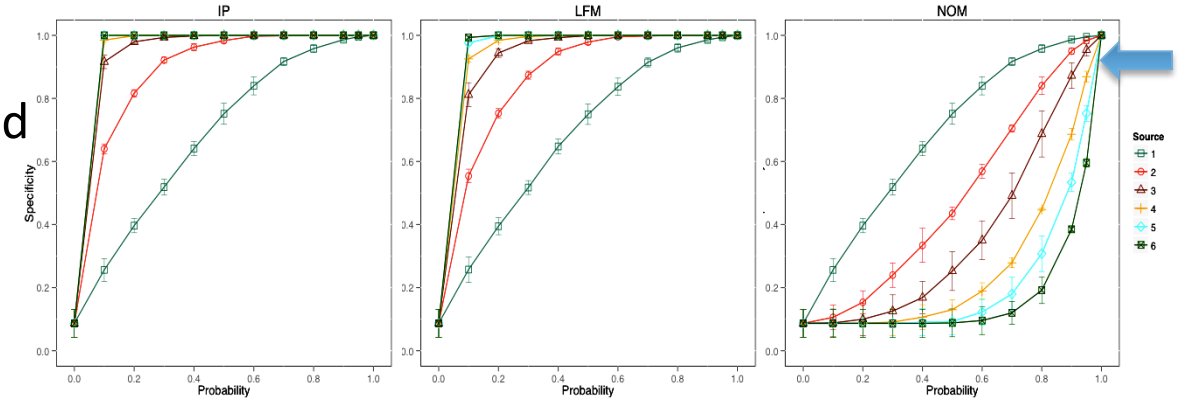


Simulations II

Number of Sources



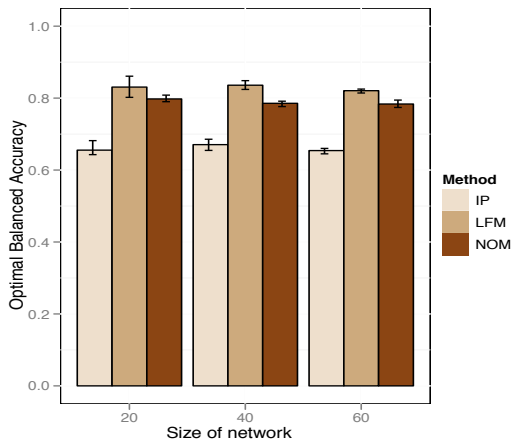


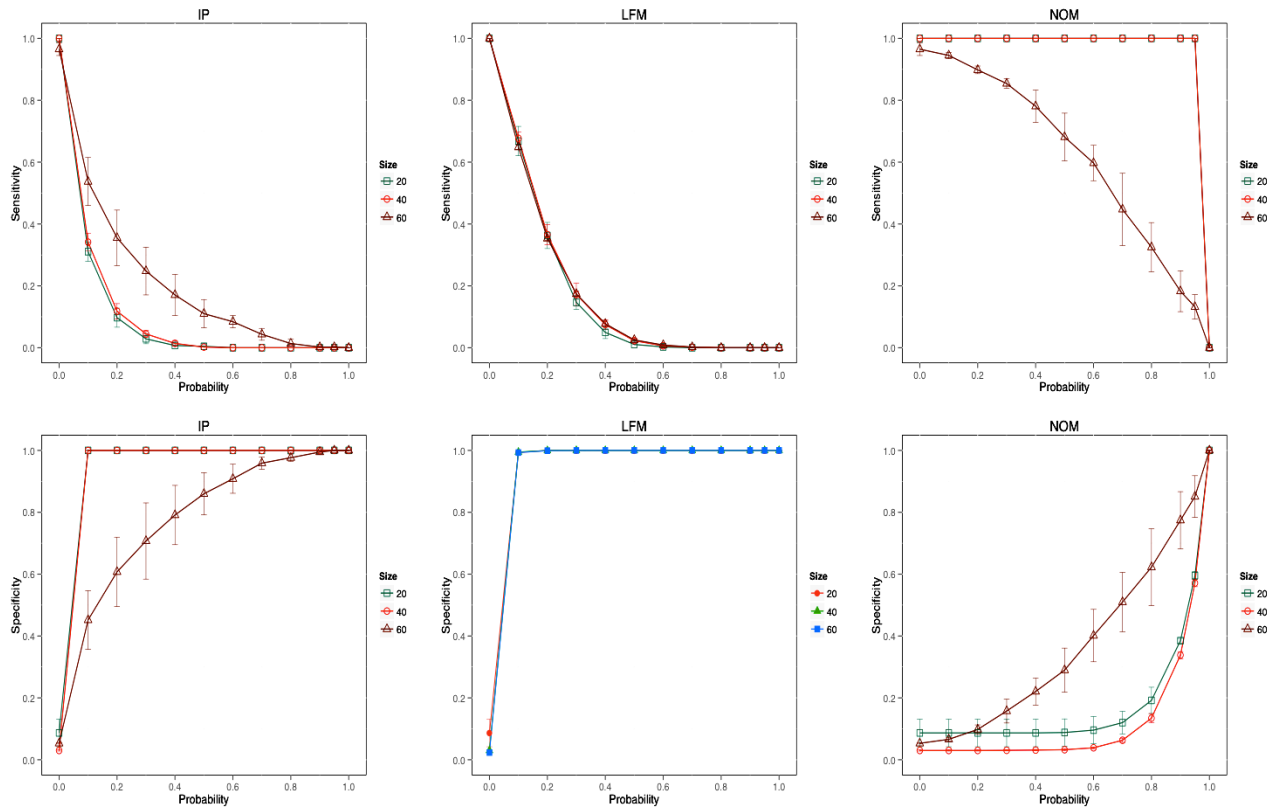


- The IP method performs worse compared to LFM and NOM with the inclusion of more source
- LFM and NOM perform better with it

Simulations II

Network Size





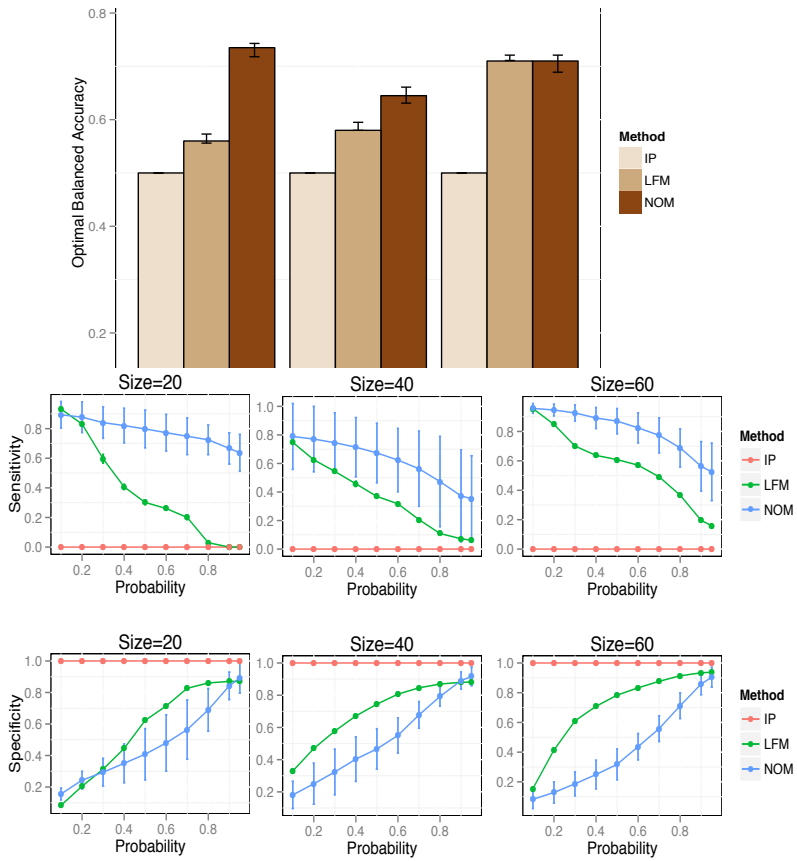
- LFM performed better even for larger networks
- But, NOM methods are faster compared to LFM

Simulations III

Measuring for KEGG Pathways

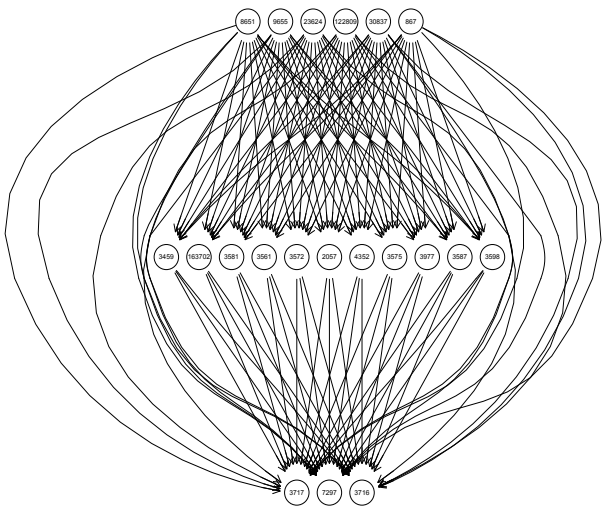
- Reconstruction of KEGG pathways only from prior knowledge derived from real information sources
- Influence of
  - Lower values of information
  - Missing information
- This leads to
  - Sensitivity close to 0
  - Specificity=1

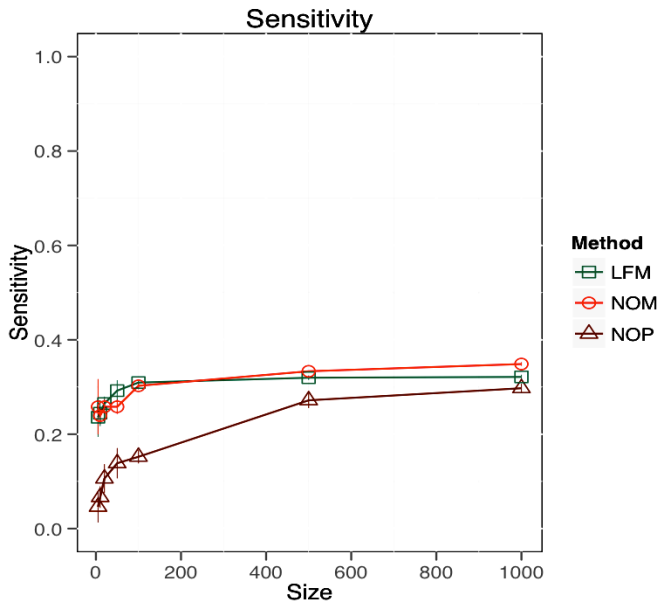
  even at very low probability cutoffs for IP method

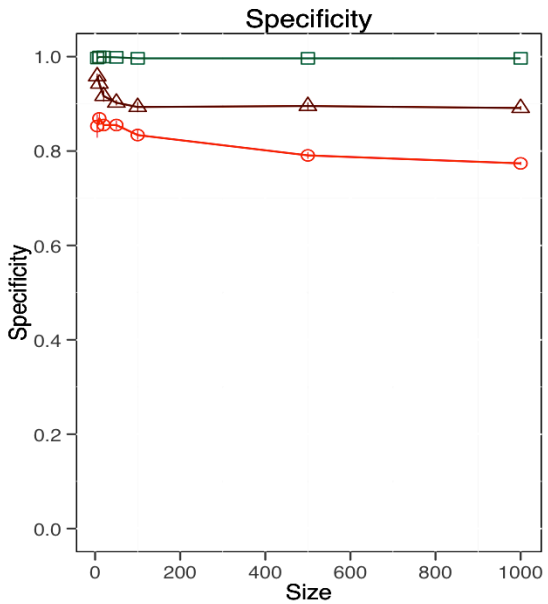


Application I

Reconstructing KEGG acyclic sub-graphs via Bayesian Networks



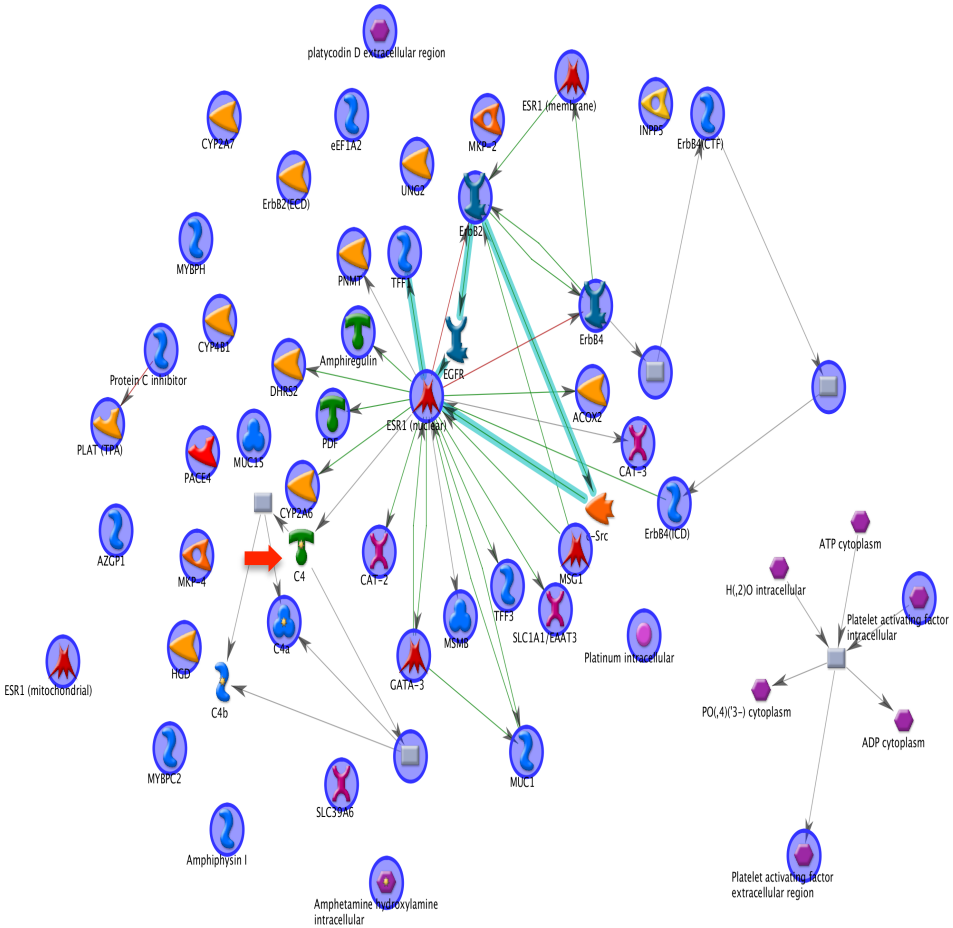


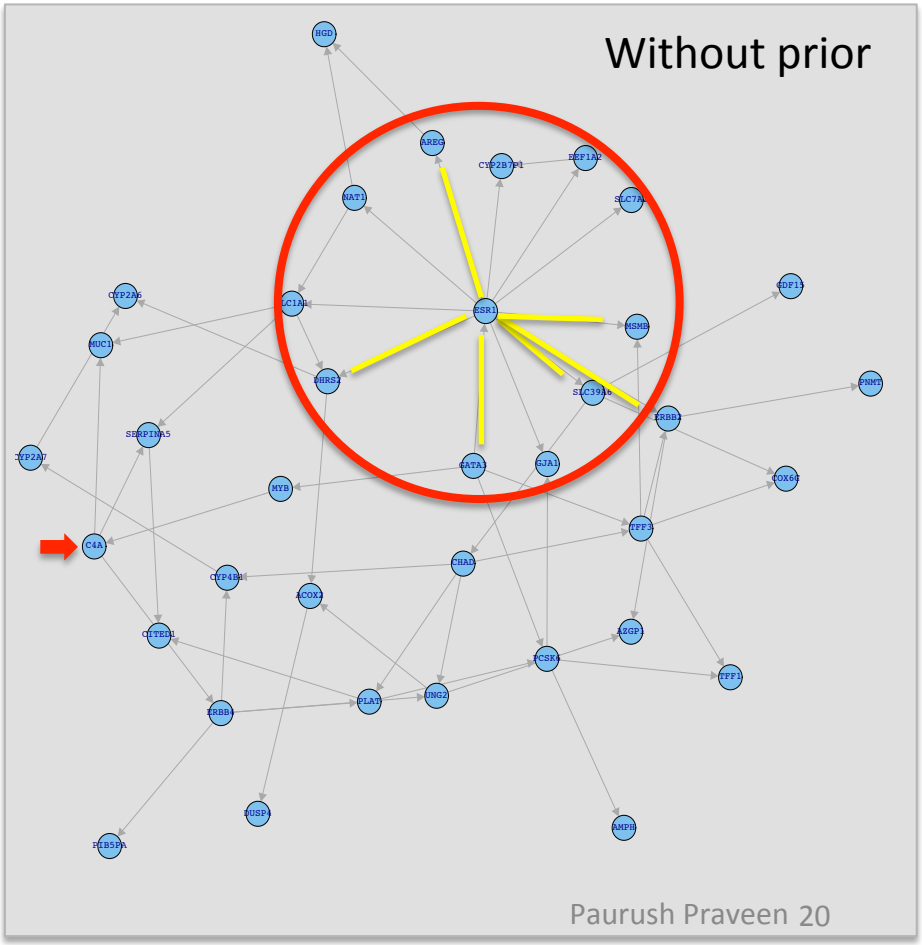


- Even for very low sample size the reconstruction with prior could achieve higher sensitivity
- The result was restricted in terms of maximum number of parents for computational feasibility.
- The LFM method maintained the highest specificity at the same time.

Application II

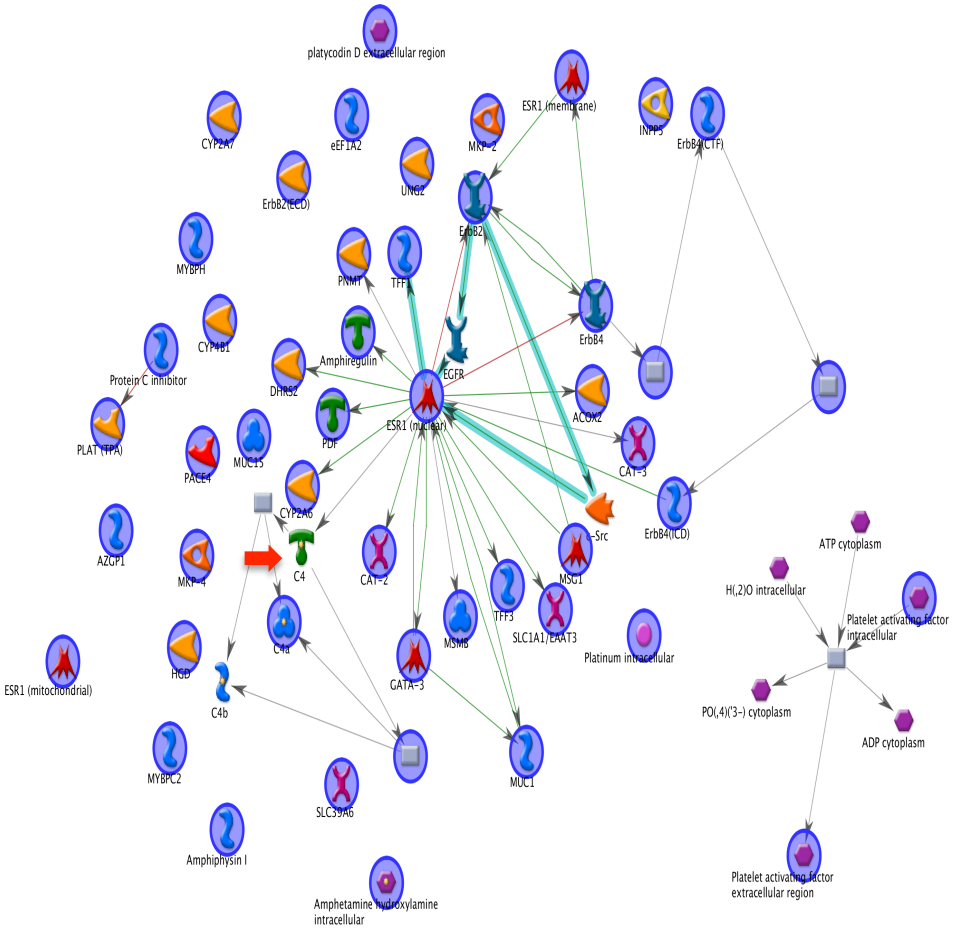
Bayesian Network for Breast Cancer Data from *van't Veer et.al (2002).*

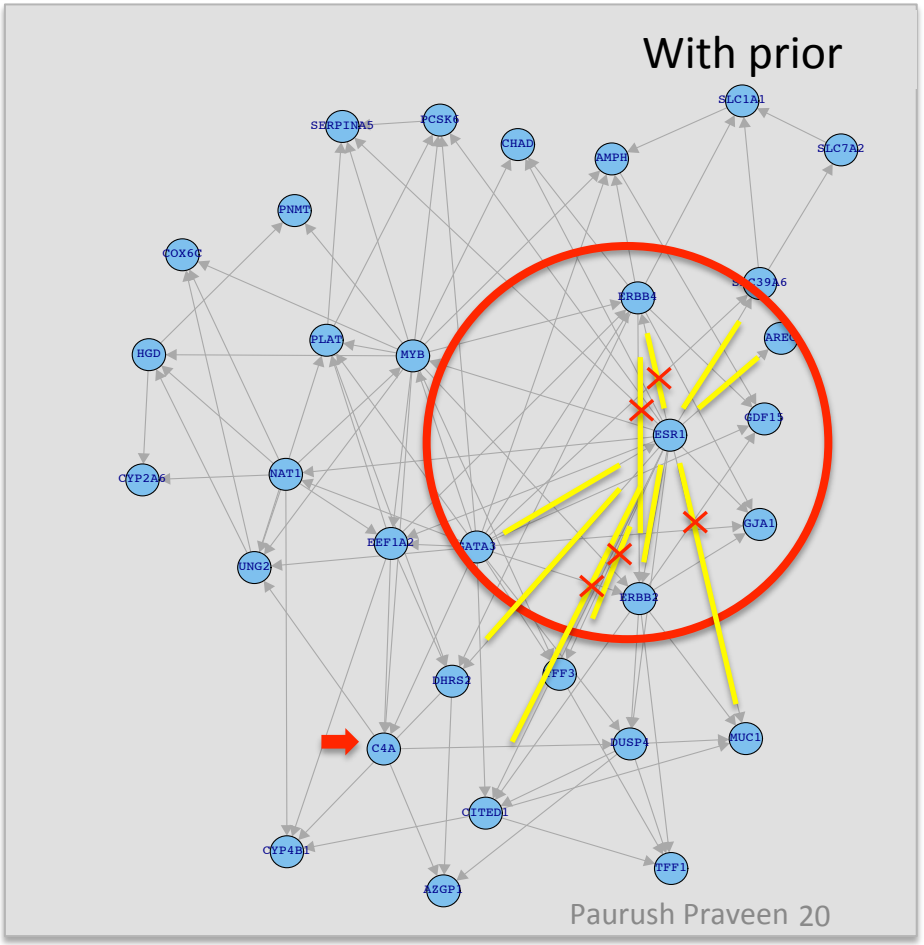




Application II

Bayesian Network for Breast Cancer Data from *van't Veer et.al (2002).*





Summary

- Two approaches for probabilistic knowledge integration for network inference from data
- Can use heterogeneous biological knowledge
- Flexible with the number of sources
- Can handle missing information
- Good agreement with known networks
- Improvement in network reconstruction accuracy using Bayesian network

Acknowledgements

- Prof. Dr. Holger Fröhlich
- Group Members at B-IT, Algorithmic Bioinformatics

- *Funding:* The work was supported in part by B-IT Research School grant.

References

1. **Mukherjee, S. & Speed, T. P. (2008) Network inference using informative priors Proceedings of the National Academy of Sciences, 105, 14313-14318**
2. **Fröhlich, H.; Fellman, M.; Sultman, H.; Poustka, A. & Beissbarth, T. (2007) Large scale statistical inference of signaling pathways from RNAi and microarray data** BMC Bioinformatics**, 2007, 8**
3. **Andreas Schlicker, Francisco S Domingues, Jörg Rahnenfhrer, Thomas Lengauer (2006) A new measure for functional similarity of gene products based on Gene Ontology.** BMC bioinformatics **7 (1) p. 302**
4. **Fröhlich, H.; Speer, N.; Poustka, A. & Beissbarth, T. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products** BMC Bioinformatics**, 2007, 8, 166.**
5. **Weile, J.; James, K.; Hallinan, J.; Cockell, S. J.; Lord, P.; Wipat, A. & Wilkinson, D. (2012) Bayesian integration of networks without gold standards** Bioinformatics
6. **Pearl, J. (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference.** Morgan Kaufmann Publishers Inc, San Francisco **1 edition.**
7. **Gao, S. &Wang, X (2011) Quantitative utilization of prior biological knowledge in the Bayesian network modeling of gene expression data** BMC Bioinformatics**, 2011, 12, 3592**

Thank-you

- Please visit the poster at MSLB-2012
- Please visit the poster (I-23) at ECCB-2012