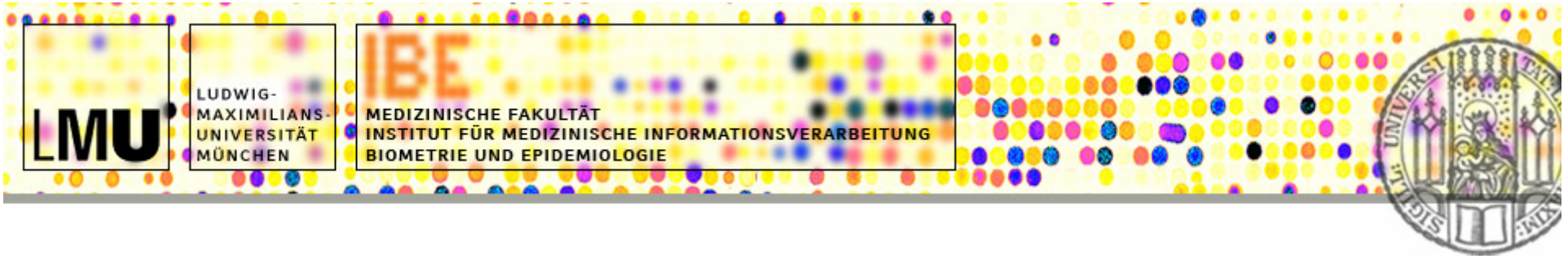


Biological aspects for the validation of estimated gene interaction networks from microarray data

Ulrich Mansmann, Vindi Jurinovic
ulrich.mansmann@lmu.de
jurinovic@ibe.med.uni-muenchen.de
IBE, LMU München

Outline

- Introduction
- Inferring networks
- Statistical validation of network algorithms
- What does microarray data quantify?
- Scale free networks
- Confounding by excluded neighbors
- The static image of dynamic processes
- MYC – Translocation
- Summary

Introduction

- Many types of biological networks exist.
- Networks are a first approach to a systems view of molecular processes within a cell.
- Few such networks are known in anything approaching their complete structure.
- Methods using high-throughput data for inference of regulatory networks rely on searching for patterns of partial correlation or conditional probabilities.
- Algorithms are designed to infer the topology of any network where the change in state of one node can affect the state of other nodes.
- Transcriptional regulatory networks, signal transduction networks, metabolite networks.
- Protein-protein interaction networks are also under very active study. Reconstruction of these networks does not use correlation-based inference.

Transcriptional regulatory networks

- Genes are the nodes.
- A gene serves as the source of a direct regulatory edge to a target gene by producing an RNA or protein molecule that functions as a transcriptional activator or inhibitor of the target gene.
- Computational algorithms used to infer the topology take as primary input the data from a set of microarray runs measuring the mRNA expression levels of the genes under consideration for inclusion in the network.
- in general, the results of the inference procedures are undirected graphs.
- Maathuis, Kalisch & Bühlmann propose a strategy to predict causal effects in large-scale systems from observational data
  (2009, Annals of Statistics 37, 3133-3164)

Estimation of graph topology

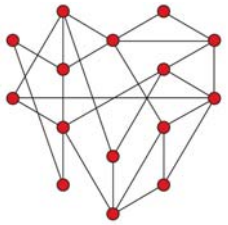
- Schäfer J, Strimmer K (2005) *A shrinkage approach to large-scale covariance estimation and implications for functional genomics*. SAGMB. 4: 32
- Meinshausen N, Bühlmann P (2006) *High dimensional graphs and variable selection with the lasso*. Annals of Statistics, 34, 1436-1462
- Kalisch M, Bühlmann P (2007) *Estimating high dimensional directed acyclic graphs with the PC-Algorithm*. Journal of Machine Learning Research, 8: 613-636
- Banerjee O, El Ghaoui L, d'Aspremont A (2008) *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data.* Journal of Machine Learning Theory, 9: 485-516.
- Friedman J, Hastie T, Tibshirani R (2008) *Sparse inverse covariance estimation with the graphical lasso*. Biostatistics. 9: 432-441

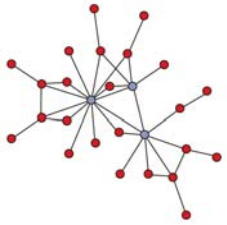
Validation by simulation

- Simulate data from a directed acyclic graph (DAG), compare the estimated topology with the topology of the moralized DAG.
- Create random networks:

  The *Erdös–Rényi (ER) model* of a random network14 starts with N nodes and connects each pair of nodes with probability p, which creates a graph with approximately p·N·(N–1)/2 randomly placed links. The node degrees follow a Poisson distribution.

  Scale-free networks are characterized by a power-law degree distribution; the probability that a node has k links follows $P(k) \sim k^{-\gamma}$, where γ is the degree exponent.



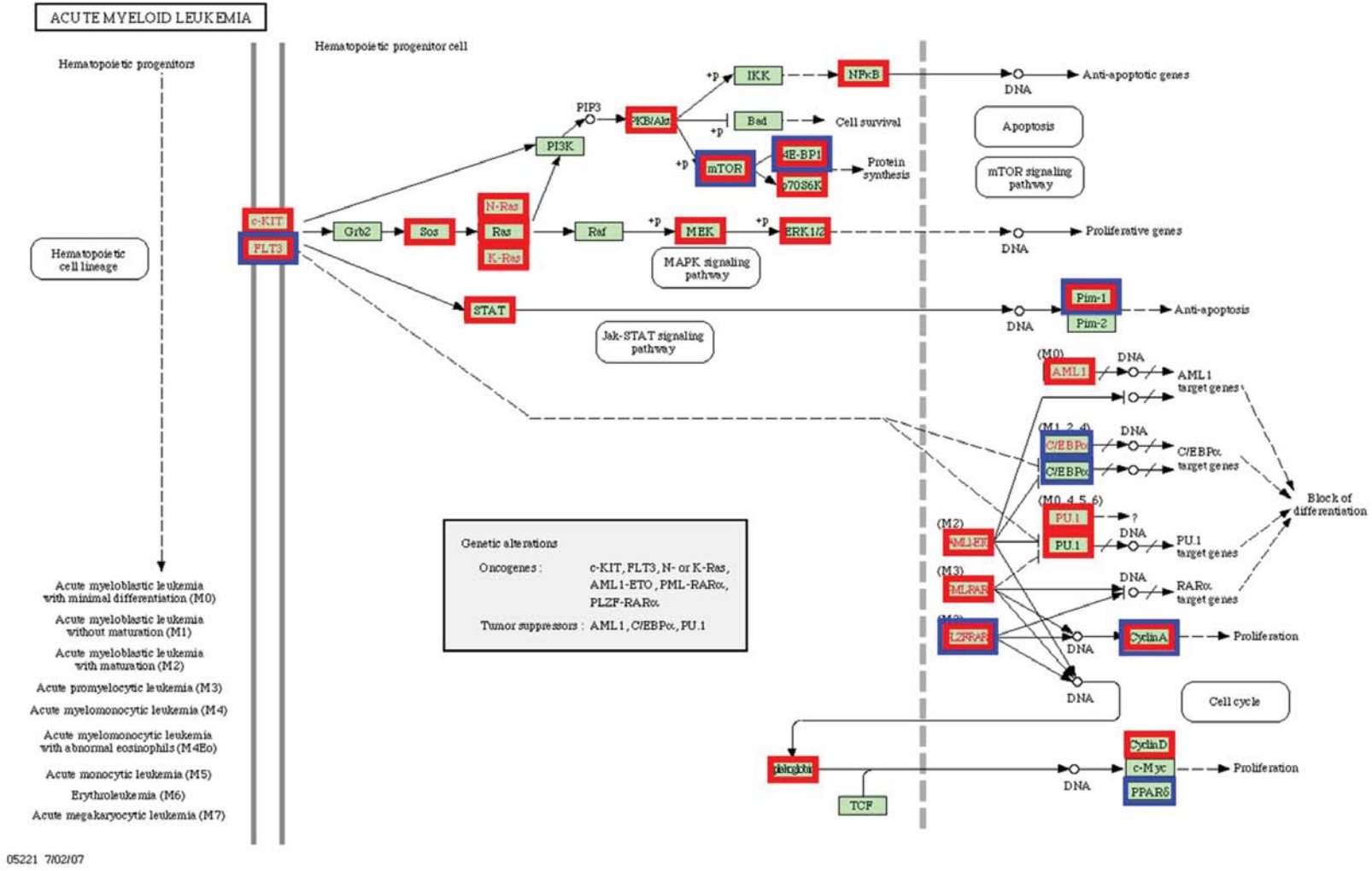


How to find a distribution which fits a pre-specified topology?

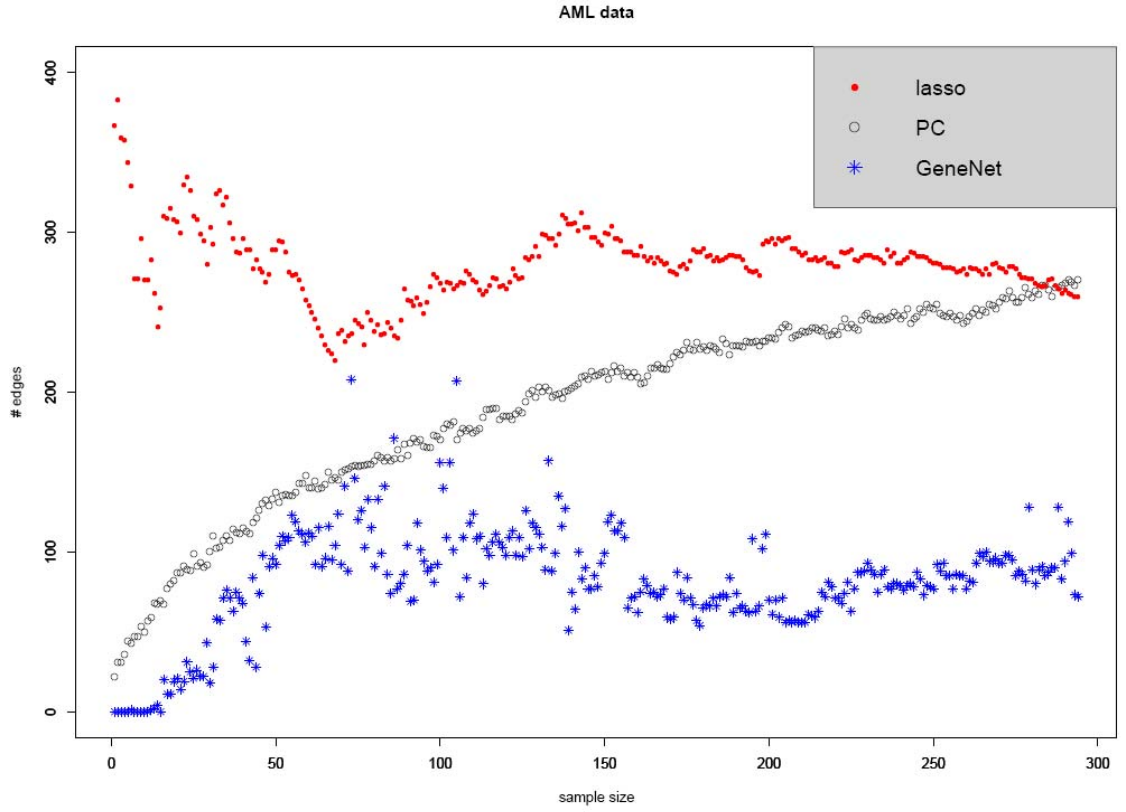
Validation by simulation

Sensitivity	percentage of edges in the original graph which are also present in the inferred graph
Specificity	percentage of missing edges in the original graph which are also not present in the inferred graph
PPV	percentage of edges in the inferred graph which are also present in the original graph
NPV	percentage of missing edges in the inferred graph which are also not present in the original graph
SHD	Structural Hamming distance between the inferred and the original graph: number of edges which have to be changed to transform one graph into the other

Validation with observed data

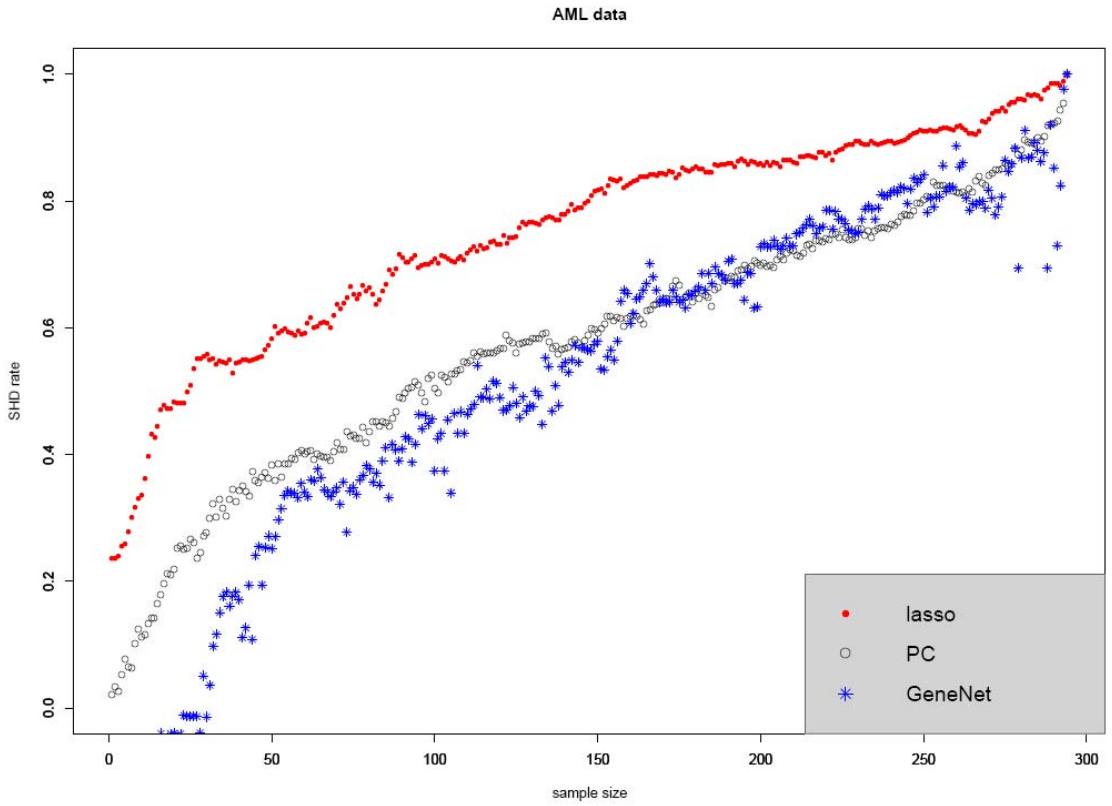


Validation with observed data

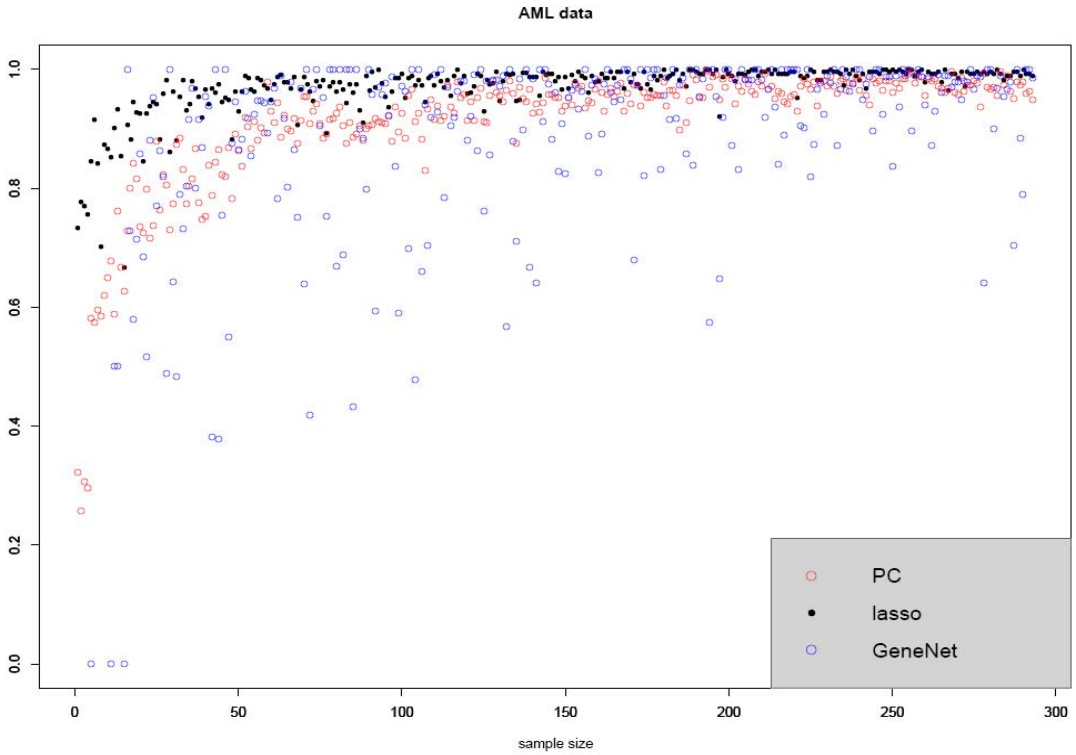


Validation with observed data

Rate = $1\text{-SHD}/(\#E_1+\#E_2)$



Validation with observed data



Percentage of edges of the graph calculated for a smaller subsample in the graph calculated from the full size.

What does microarray data quantify?

- Cumulative value of observed gene expression in a large ensemble of cells: cross-sectional observation.
- Time average of a dynamic process – assuming some ergodicity arguments
- Observed steady-state of a dynamic process: heuristic argument for reproducible observations

Averaged dynamics

The (high-dimensional) multivariate process $(X_t)_{t\varepsilon[0,T]}$ describes the dynamic of all entities (gene expression, protein concentration, …) within the cell over a time period of length T.

If a large amount of cells is assayed, we can assume that all possible time states which can be taken by the dynamic of the process are observed within the ensemble.

Under mild assumptions, the average of all measurements over the cells is identical to the average over time of $(X_t)_{t\varepsilon[0,T]}$

$$W = \frac{1}{T}\int_0^T X_t \, dt$$

Averaged dynamics

$$W = \frac{1}{T}\int_0^T X_t \, dt$$

Given that W is observed in the tissue of several individuals, it is possible to estimate the conditional correlation structure of W by a graphical model and to come up with a conditional correlation graph.

What does W represent?

Internal structure of noise

$$X_t = F(t) + \varepsilon_t \qquad W = \frac{1}{T}\int_0^T X_t \, dt \qquad E[\varepsilon_t] = 0 \qquad E[\varepsilon_t \otimes \varepsilon_s] = \Psi_{s-t}$$

F(t) is the deterministic process which describes the dynamics of all cell constituent parts.

The stationary mean zero spatio-temporal process ε_t quantifies the noise and has cross-covariance Ψ_{s-t}.

The noise may be heterogeneous over the cell and homogeneous within local structures. In different regions, noise may also depend on different time scales.

Both introduces a complex spatio-temporal covariance structure for ε.

This structure influences the measurement of the complex biological signals. The covariance of W is defined by the complex spatio-temporal dependency structure of the noise process and not by dependencies in the underlying biological process.

$$Cov(W) = \frac{1}{T^2}\int_0^T E[\varepsilon_t \otimes \varepsilon_s] \, dt \, ds$$

Driven by transcription factors

$$X_t = f(t) + \varepsilon_t^X$$
$$Y_t = a \cdot X_{t-\delta} + \varepsilon_t^Y$$
$$Z_t = b \cdot Y_{t-\Delta} + \varepsilon_t^Z$$

$$\begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix} = \begin{bmatrix} f(t) \\ a \cdot f(t-\delta) \\ a \cdot b \cdot f(t-\delta-\Delta) \end{bmatrix} + \begin{bmatrix} \varepsilon_t^X \\ a \cdot \varepsilon_{t-\delta}^X + \varepsilon_t^Y \\ b \cdot a \cdot \varepsilon_{t-\delta-\Delta}^X + b \cdot \varepsilon_{t-\Delta}^Y + \varepsilon_t^Z \end{bmatrix}$$

Assumption: Δ and δ << T and ε is white independent noise

$$\Sigma = \mathrm{Cov}\left(\begin{bmatrix} W^X \\ W^Y \\ W^Z \end{bmatrix}\right) = \begin{bmatrix} 1 & a & a \cdot b \\ a & 1+a^2 & a^2 \cdot (1+b) \\ a \cdot b & a^2 \cdot (1+b) & a^2 \cdot (1+b^2)+1 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1+a^2 & -a & 0 \\ -a & 1+b^2 & -b \\ 0 & -b & 1 \end{bmatrix}$$

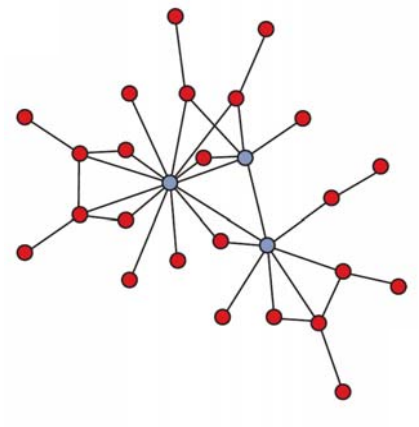
The result is less simple if the noise process has an autoregressive structure.

Scale-free networks

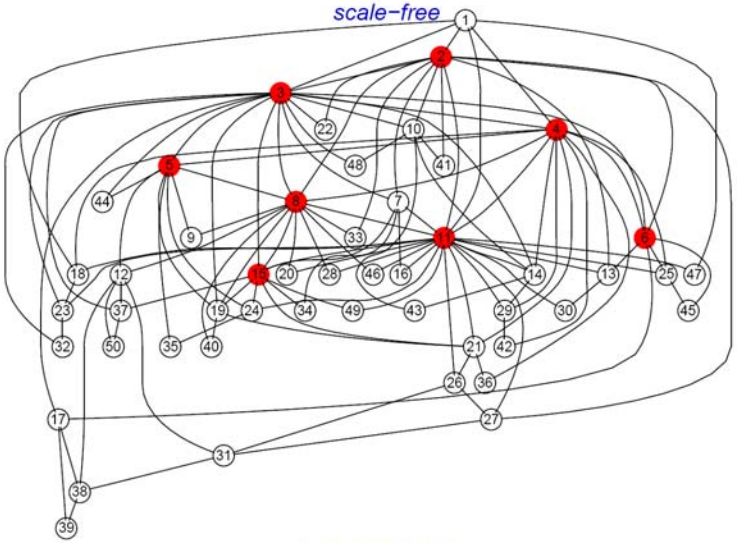
Albert-László Barabás & Zoltán N. Oltvai (2004) Network Biology: *Understanding the cell's functional organization*, Nature reviews Genetic, 5:101-114

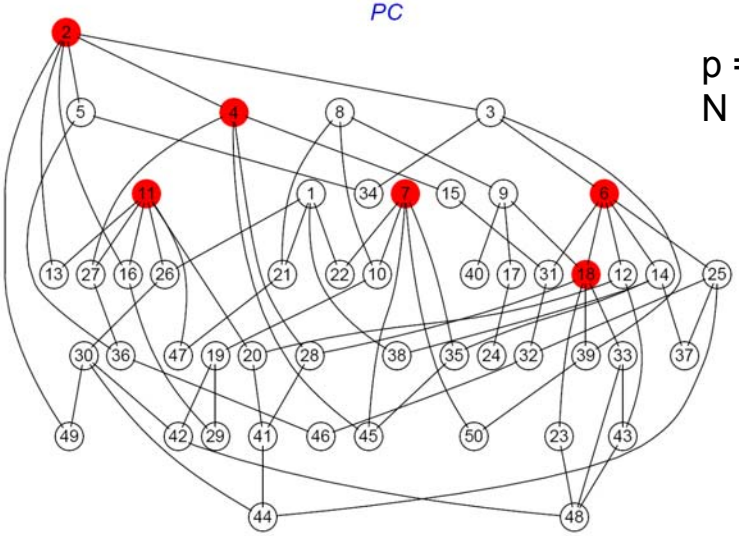
Examples of scale-free organization include genetic regulatory networks, in which the nodes are individual genes and the links are derived from the expression correlations that are based on microarray data.

The distribution that captures how many different genes a transcription factor interacts with follows a power law, which is a signature of a scale-free network. This indicates that most transcription factors regulate only a few genes, but a few general transcription factors interact with many genes.

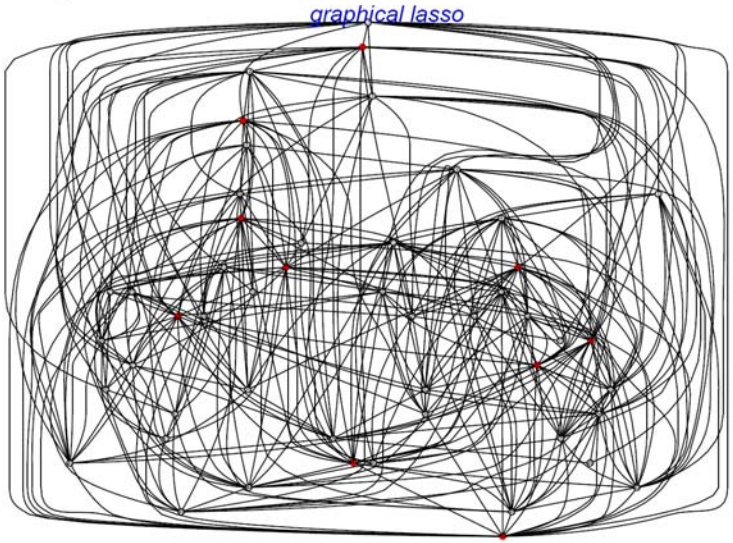


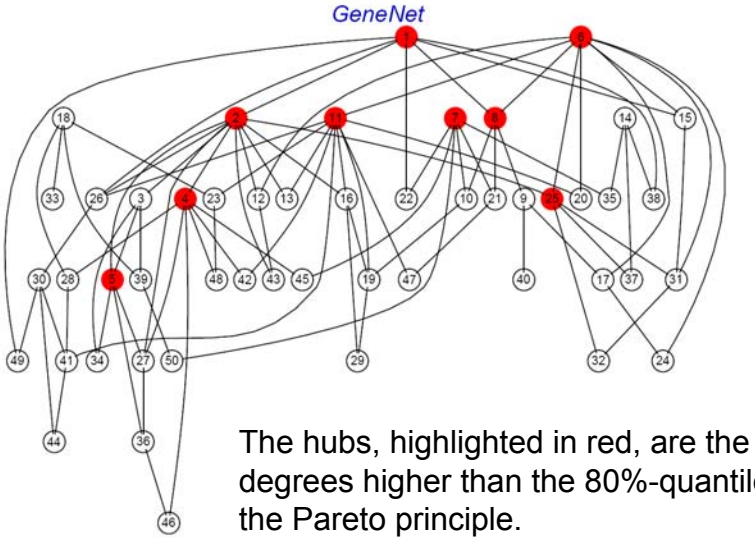
Scale-free networks





p = 50
N = 500





The hubs, highlighted in red, are the nodes with degrees higher than the 80%-quantile, following the Pareto principle.

Confounding neighbors

W = (U,V) U observed components
V unobserved components

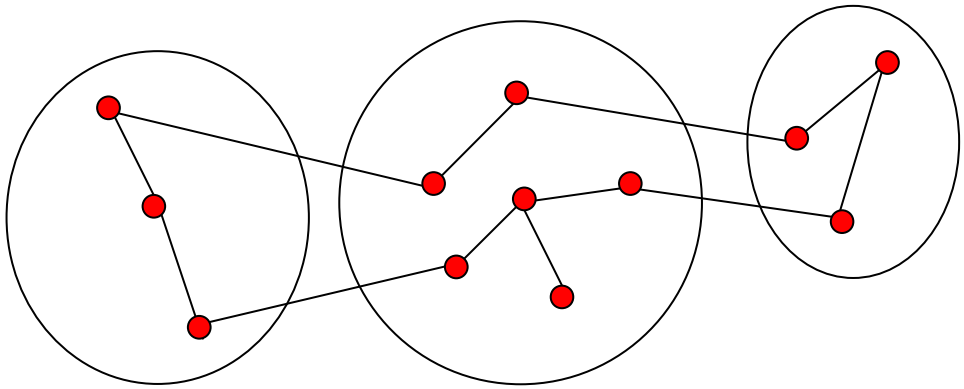
$Q = Cov(W)^{-1}$

$$Q = \begin{pmatrix} Q_{UU} & Q_{UV} \\ Q_{VU} & Q_{VV} \end{pmatrix}$$

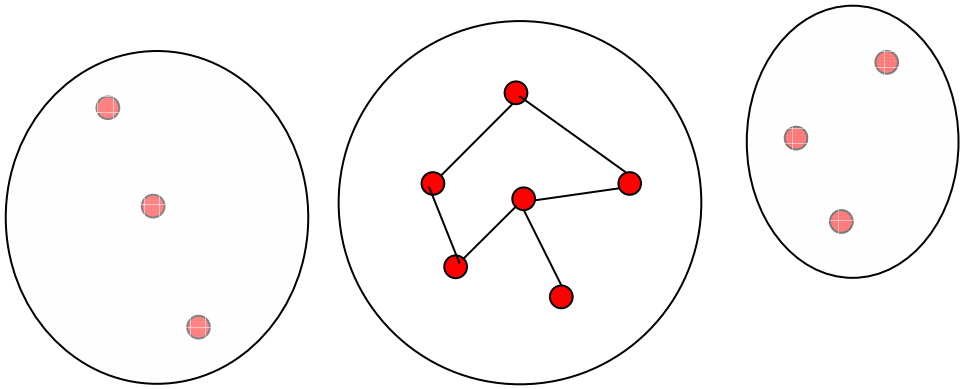
$$Q_{UU}^{marg} = Q_{UU} + Q_{UV} \cdot Q_{VV}^{-1} \cdot Q_{VU}$$

Consider a transcription factor that regulates the expression of a gene G_1 and belongs to the unobserved components V. The concentration of a transcription factor may be regulated by some other protein which is also an unobserved component in V. This protein is regulated by the transcriptional products of gene G_2. The conditional correlation structure of both proteins is an element of Q_{VV} while the interaction of the transcription factor with G_1 and the protein regulation by gene G_2 are represented by elements of QUV. This may imply a non-zero element in Q_{UU}^{marg} without the need for direct interaction of G_1 and G_2 within the pathway.

Confounding neighbors

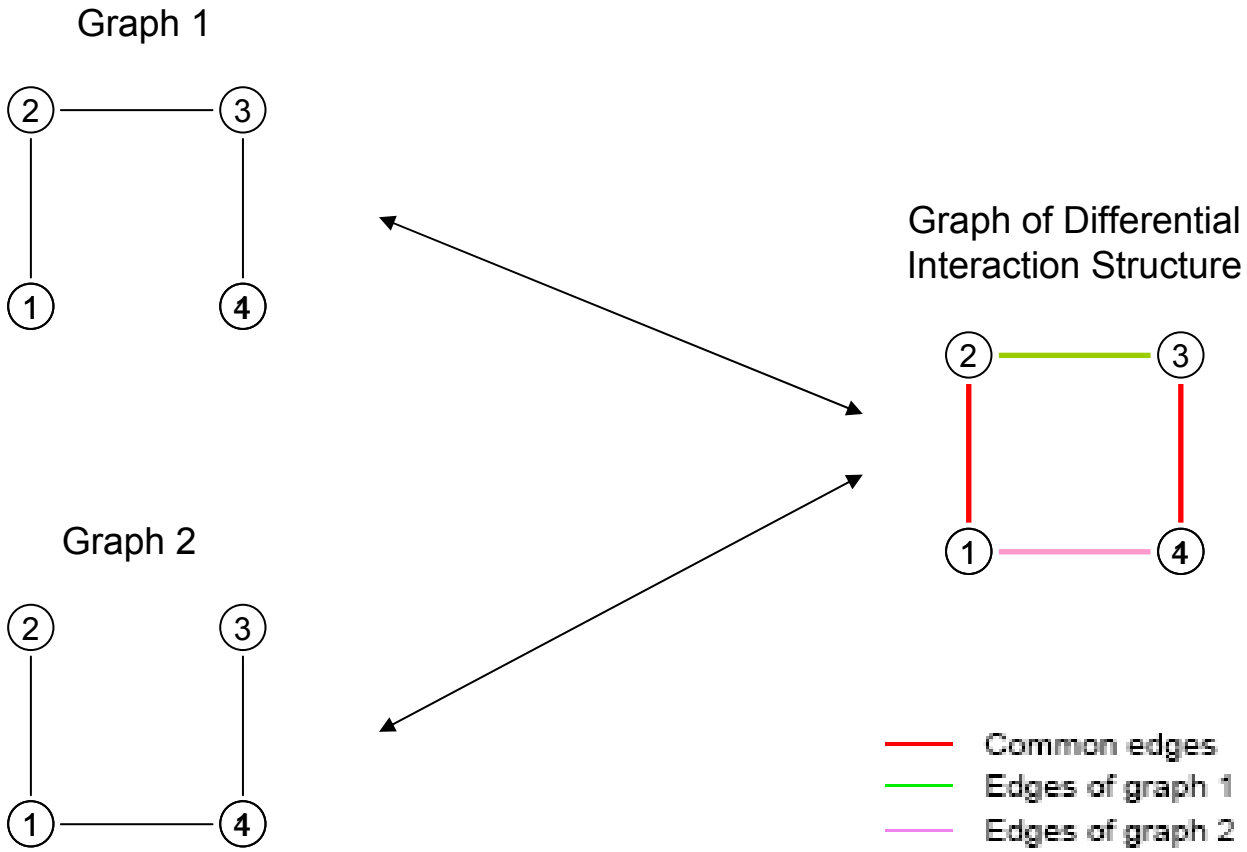


extended analysis

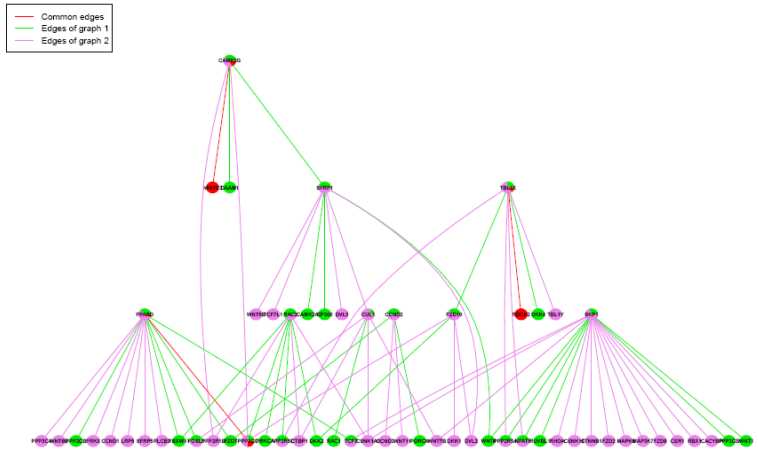


restricted analysis

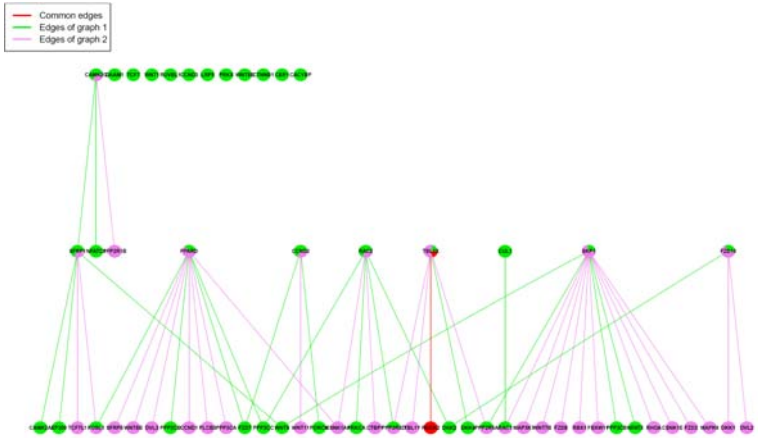
Contrasting Graphs: Differential interaction structure



Confounding neighbors



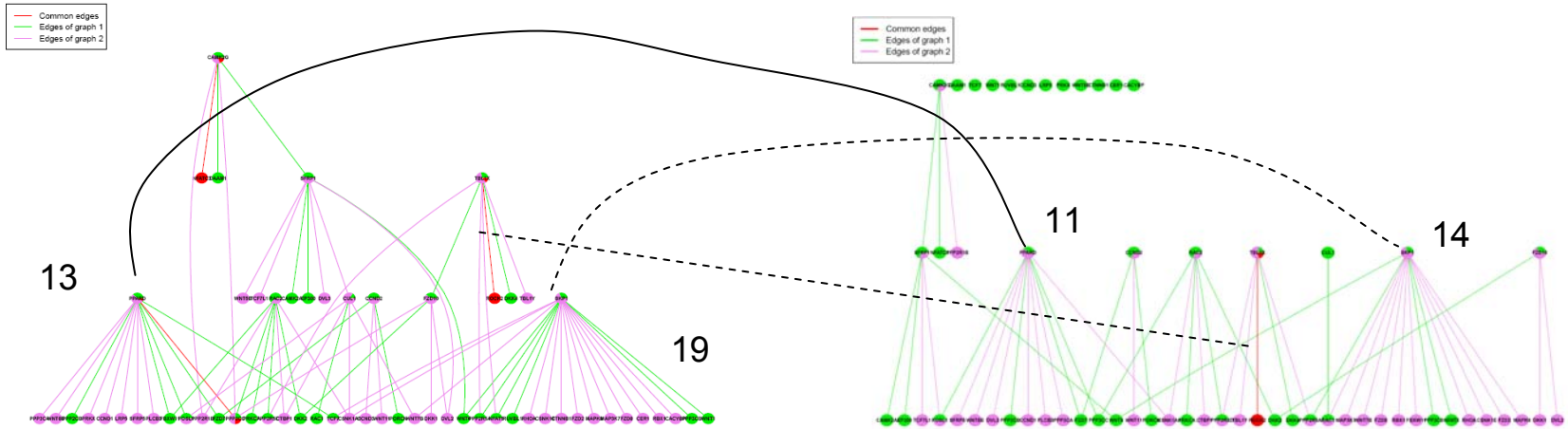
Differential interaction structure for the *wnt* pathway



Differential interaction structure for the *wnt* pathway when genes from cell-cycle are incorporated into the inference

Graph 1: Interaction estimate for IG-translocated samples
Graph 2: Interaction estimate for samples with normal karyotype

Confounding neighbors

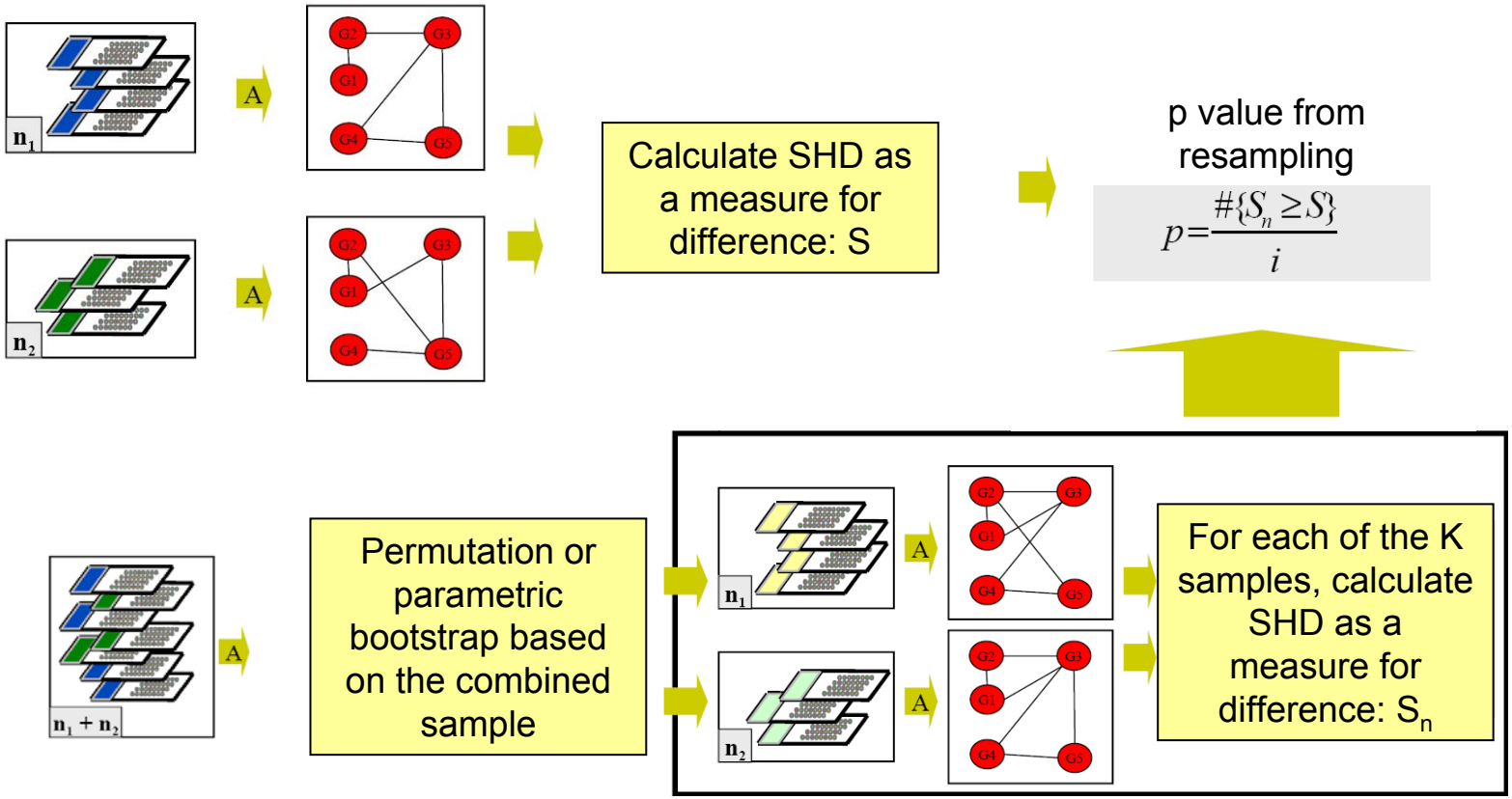


Differential interaction structure for the wnt pathway

Differential interaction structure for the wnt pathway when genes from cell-cycle are incorporated into the inference

Graph 1: Interaction estimate for IG-translocated samples
Graph 2: Interaction estimate for samples with normal karyotype

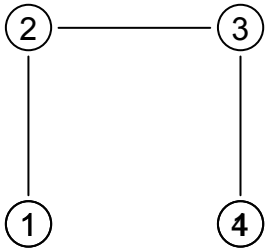
Comparing graphs: the topology oriented approach

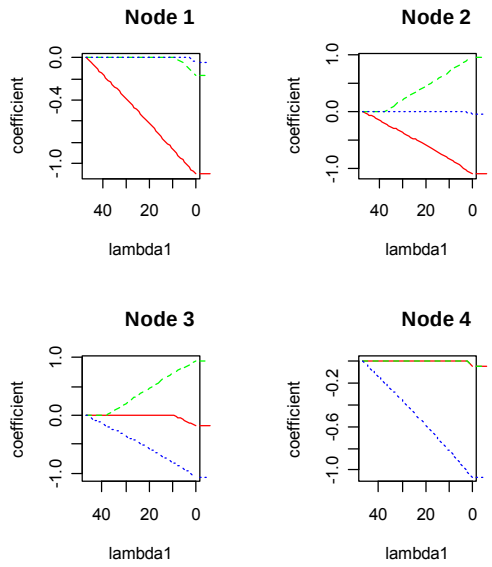


The direct approach uses an algorithm to estimate graphs.

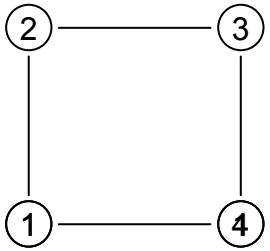
Comparing graphs: DDIS

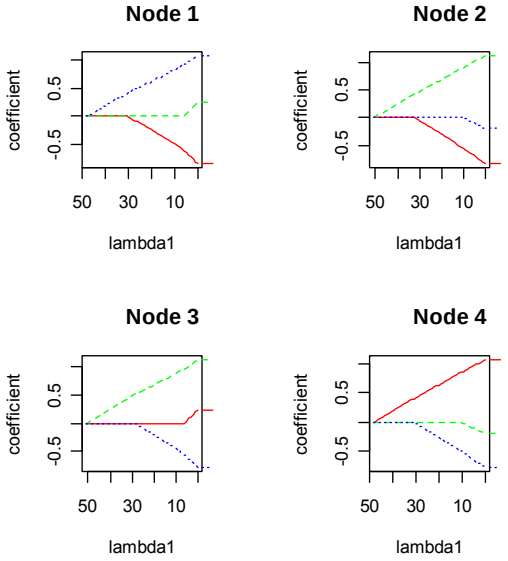
Graph 1





Graph 2





Comparing graphs: DDIS

- Describe the interaction of node i with the remaining nodes by the lasso path plot for the regression of all remaining nodes on node i.
- This avoids to make a specific choice for the penalty parameter.
- Define for two path plots created from the Lasso regression for node i under both data sets the function

  $\Phi_i(\lambda_1,\lambda_2)$ = (# of common nonzero regression coefficients of equal sign by penalties λ_1, λ_2)

- Look for node i at

$$\Psi_i = \iint_{[0,\infty[^2} \Phi_i(\lambda_1,\lambda_2)\, d\lambda_1\, d\lambda_2$$

- Define for the entire set of nodes $\Psi = \sum_{\text{nodes}} \Psi_{\text{node}}$
- Define for a subset S of nodes $\Psi_S = \sum_{\text{node} \in S} \Psi_{\text{node}}$

MYC and its translocation

The data is taken from

Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF et al. (2006) *A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling*. N Engl J Med 354: 2419–2430.

and consists of samples from 199 patients:

140 patients with normal karyotype, 59 patients with MYC-IG translocation.

Affymetrix HGU133A raw data were normalised and statistical calculations were done using Bioconductor and R software.

MYC and its translocation

The **MYC gene** produces a **transcription factor c-Myc** that controls cellular proliferation, programmed cell death and differentiation. It is a part of several important pathways and is therefore well suited to study the effect of confounding when inferring pathway specific networks.

Furthermore, the MYC gene can be translocated and placed at other loci in the DNA. One important translocation involves the immunoglobulin heavy chain gene locus (IGH) on chromosome 14.

When placed in a region of vigorous gene transcription like the IGH locus, it is overexpressed and causes uncontrolled cell proliferation.

Thus, the transcription factor c-Myc has impact on cell cycle dynamics which is of interest for its effects on the steady state data given by the microarrays.

MYC and its translocation: effect on CC

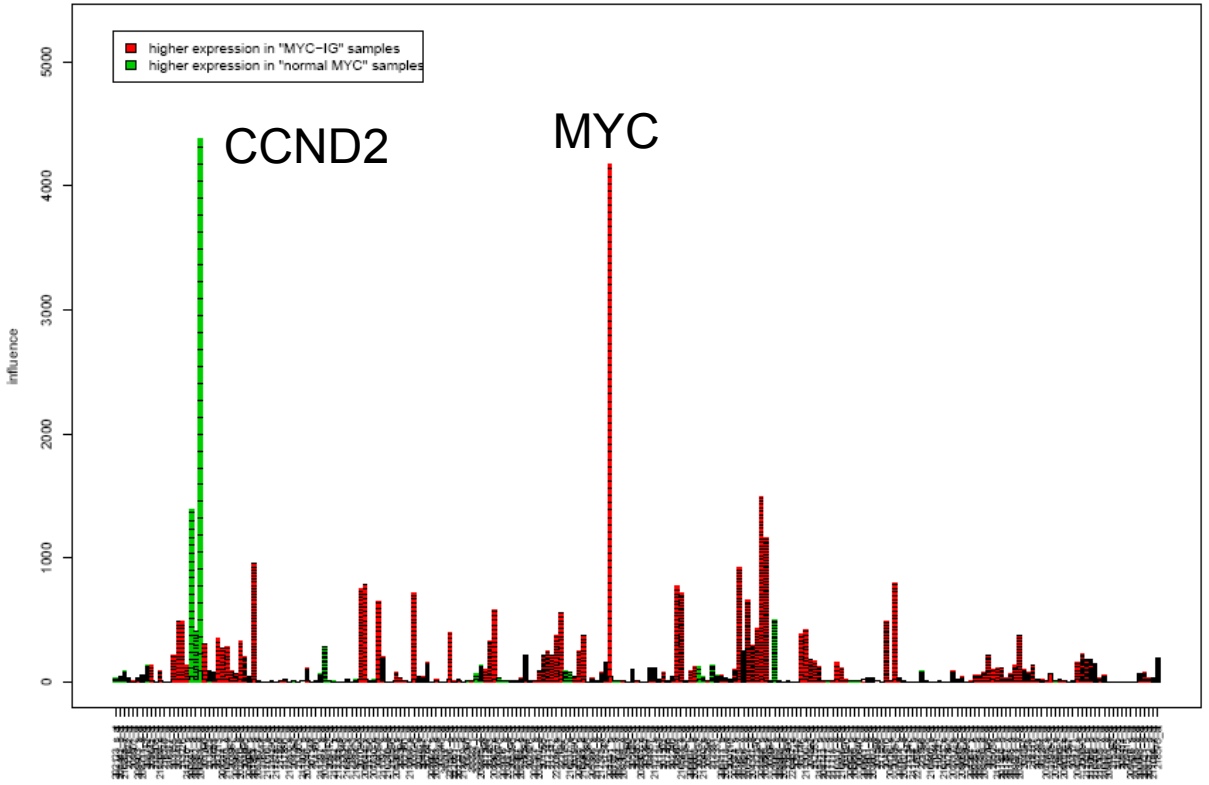
Overexpression of c-Myc largely influences the cell cycle, for example by stimulating or antagonizing the activity of different genes.

Overexpression of c-Myc in growing cells causes reduced growth requirements and shortens the G1 phase, while its underexpression leads to a lengthening of the cell cycle.

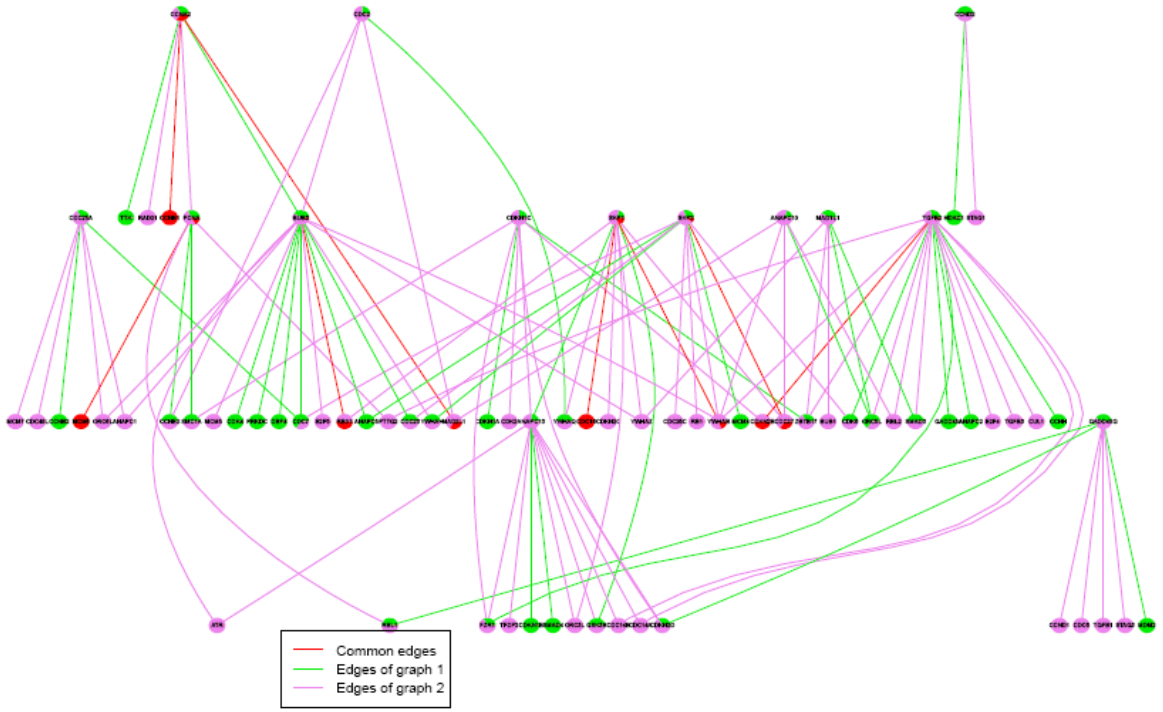
Study the *cell-cycle pathway* between patients with IG-translocated and therefore overexpressed c-Myc, and patients with normal karyotype.

Naiv hypothesis: Since the overexpression of c-Myc only stimulates the cell cycle but doesn't disrupt it, we expect its genes to be overexpressed in MYC translocated cells, but *less* disturbed in their interactions.

MYC and its translocation: effect on CC



MYC and its translocation: effect on CC



Graph 1: Interaction estimate for IG-translocated samples
Graph 2: Interaction estimate for samples with normal karyotype

MYC and its translocation: effect on CC

SHD based comparison: The resulting permutation p-value is 0.009 which we rate as evidence in a global test for differential interaction. The graph does not contain the MYC gene.

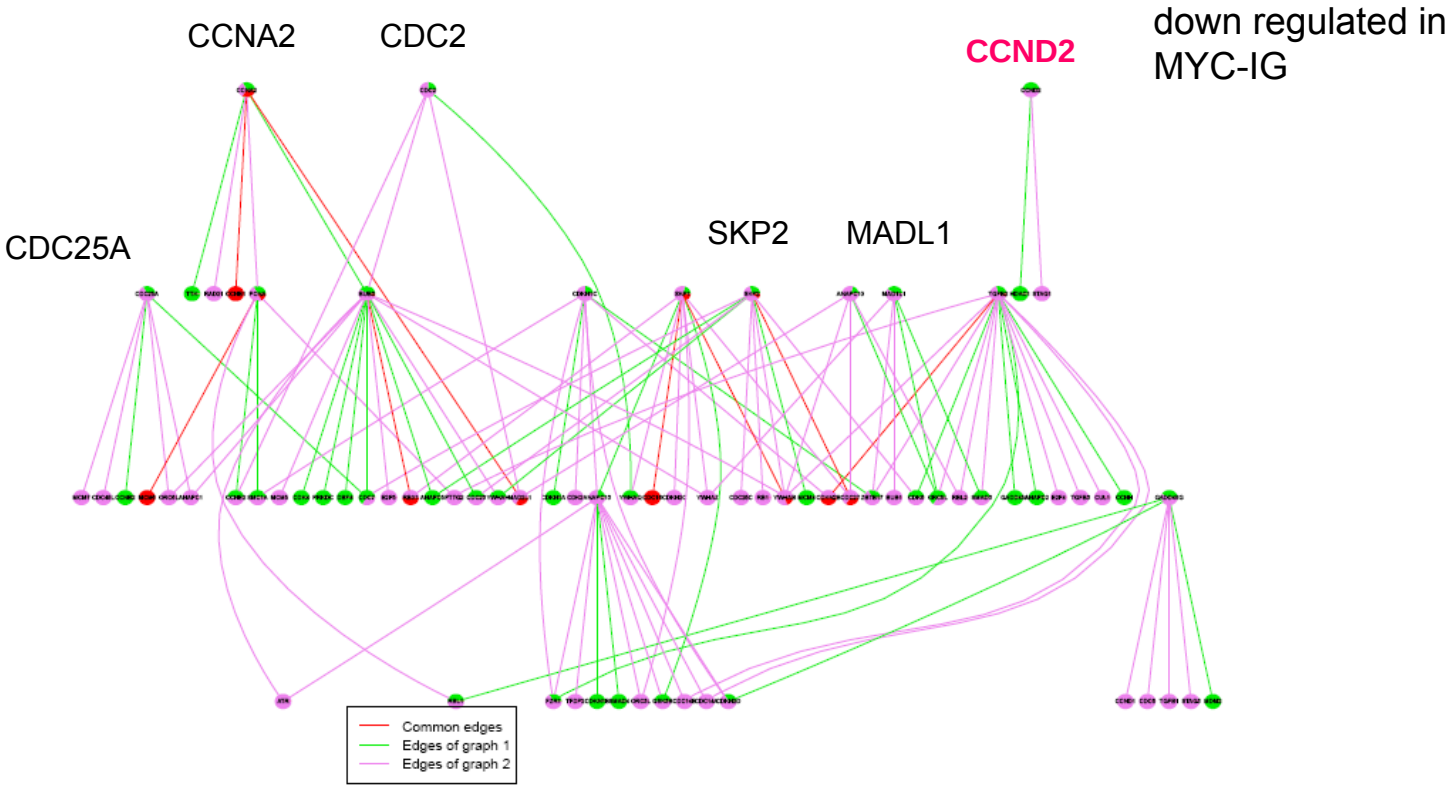
DDIS based comparison: No strong evidence for differential interaction between MYC-IG and normal samples

There are 14 nodes (probesets) with a p-value below 0.05 (representing 14 genes; **CCNA2, CCND2, CDC2, CDC25A, CDKN1C, PCNA, SKP1, SKP2, TGFB2, MAD1L1, BUB3, ANAPC10, GADD45G, ANAPC13**). Adjustment for multiple testing does not result in any significant node.

Only one of these genes is contained in the set of strongly down regulated (in MYC-IG) genes (**CCND2**) and five are contained in the set of strongly up regulated (in MYC-IG) genes (**CCNA2, CDC2, CDC25A, SKP2, MAD1L1**).

All genes detected on the unadjusted level by the DDIS method are also present in the graph comparison given in the figure as hubs.

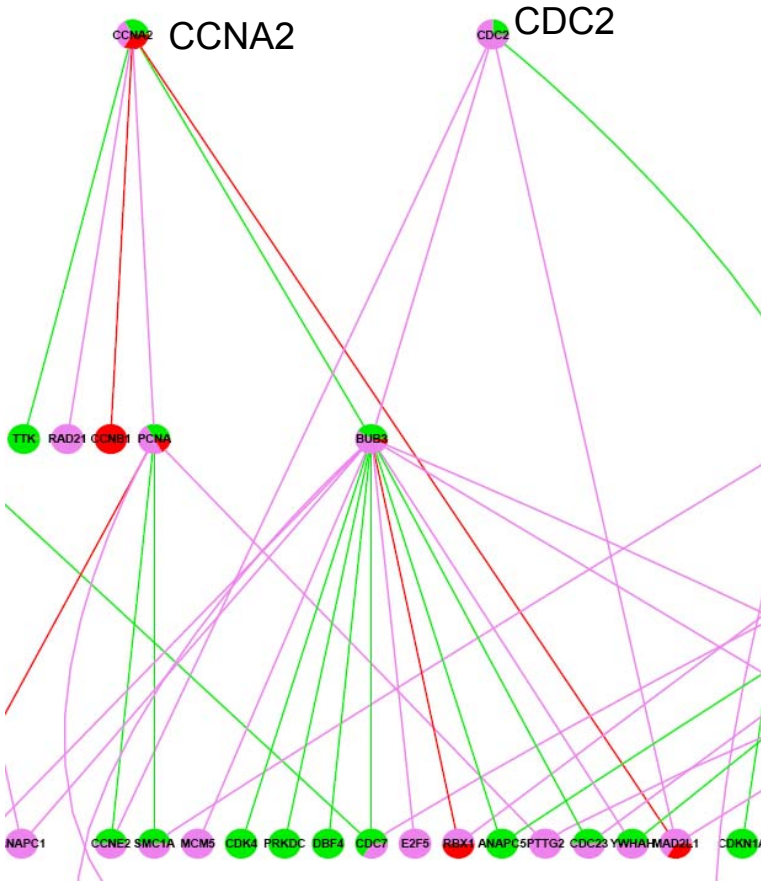
MYC and its translocation: effect on CC



Graph 1: Interaction estimate for IG-translocated samples
Graph 2: Interaction estimate for samples with normal karyotype

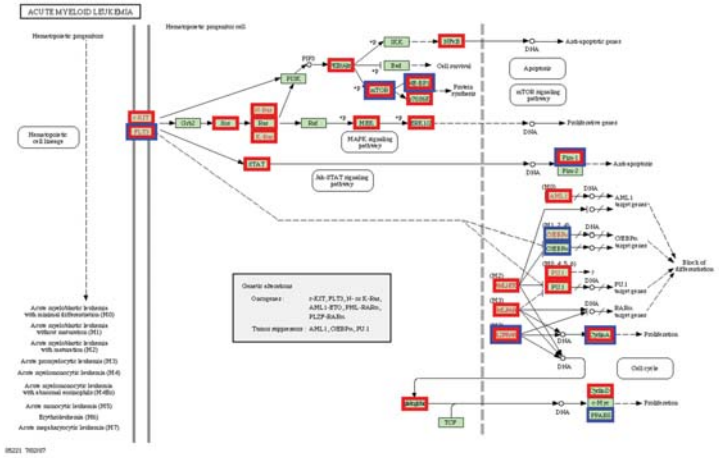
MYC and its translocation: effect on CC

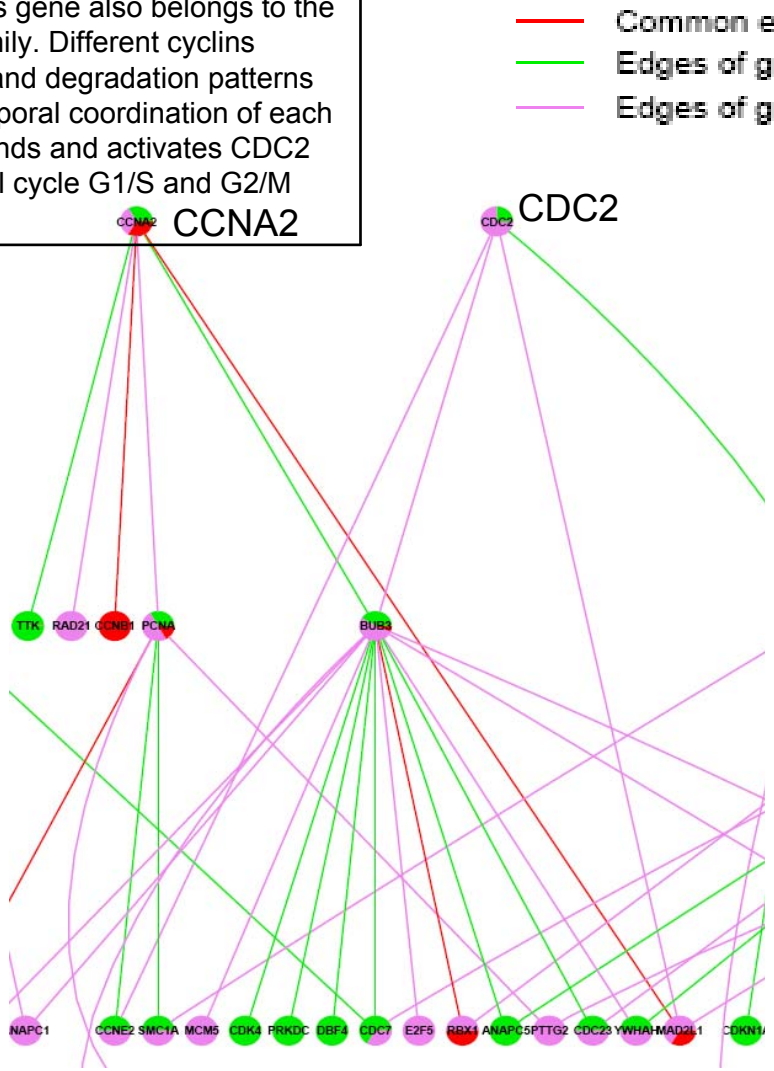
CCNA2	Cell-cycle	The protein encoded by this gene also belongs to the highly conserved cyclin family. Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of each mitotic event. This cyclin binds and activates CDC2 and thus promotes both cell cycle G1/S and G2/M transitions.



Scoring the coherence between inferred network and biological knowledge

CCNA2	Cell-cycle	The protein encoded by this gene also belongs to the highly conserved cyclin family. Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of each mitotic event. This cyclin binds and activates CDC2 and thus promotes both cell cycle G1/S and G2/M transitions.





MYC and its translocation: Interaction with IGH

The interaction of IGH gene and the MYC gene in cells with and without translocation:

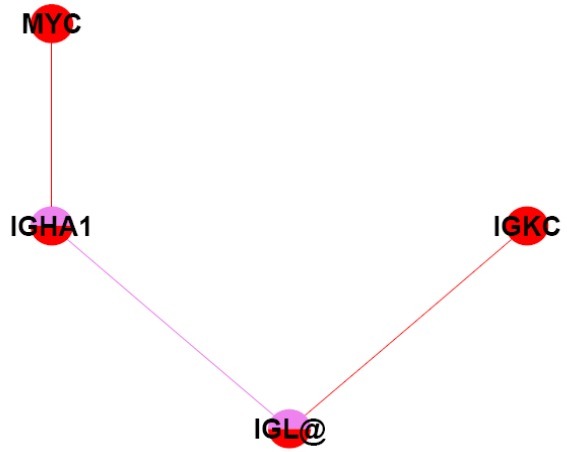
In cells with translocation MYC and IGH share parts of the same promoter.

IGH is constantly activated by its promoter in normal cells.

This causes MYC overexpression in translocated cells which react by downregulating MYC.

Since the MYC promoter is part of the IGH promoter, the reaction causes underexpression of IGH.

MYC and its translocation: Interaction with IGH



restricted inference

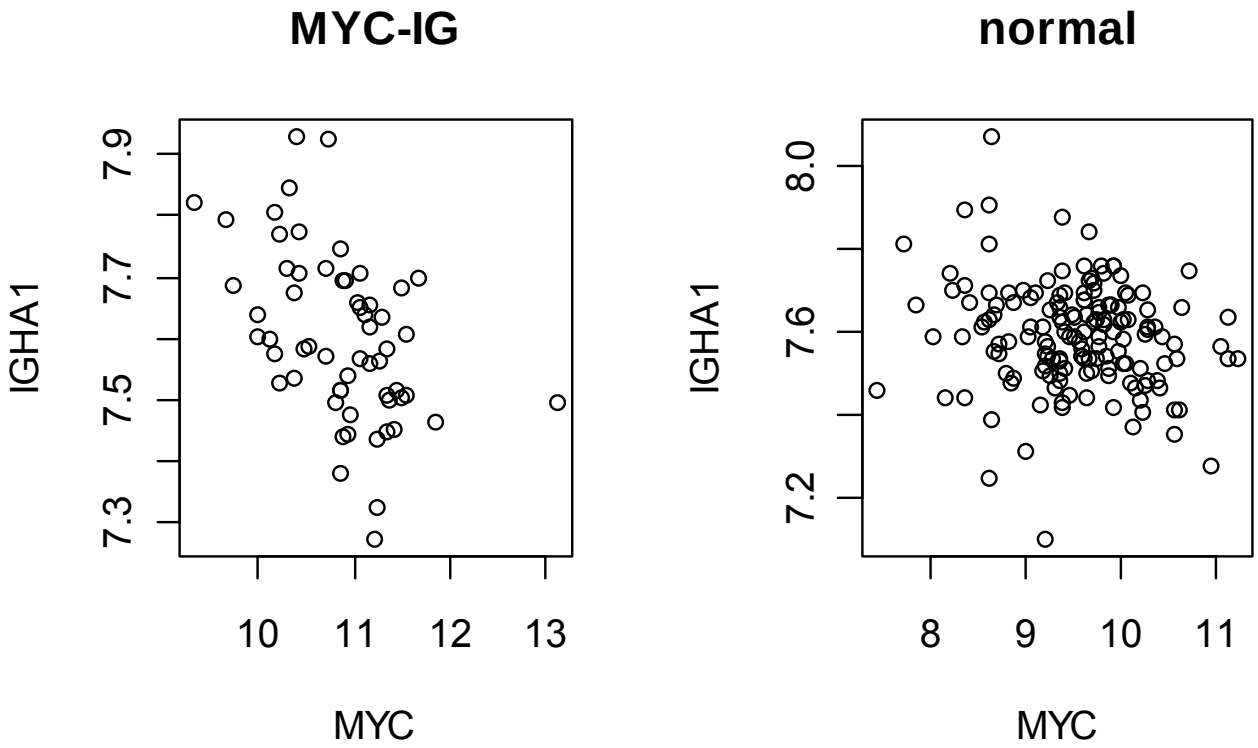
extended inference

Red edges: Direct interaction present for the translocated as well as the normal samples.

Green edges: Direct interaction present only for the translocated samples.

Purple edges: Direct interaction present only for the normal samples.

MYC and its translocation: Interaction with IGH



MYC and its translocation: Dynamics

$$X_t = \begin{bmatrix} 1 \\ X_t^1 \\ \vdots \\ X_t^m \end{bmatrix} = A \cdot \begin{bmatrix} 1 \\ X_{t-\delta}^1 \\ \vdots \\ X_{t-\delta}^m \end{bmatrix} + \begin{bmatrix} 0 \\ \varepsilon_t^1 \\ \vdots \\ \varepsilon_t^k \end{bmatrix}$$

$$X_k = A \cdot X_{k-1} + \varepsilon_k$$

$$\varepsilon_k \sim N(0, \Omega)$$

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_k^2 \end{pmatrix}$$

$$X_k = A^k \cdot X_0 + \sum_{h=1}^{k} A^{k-h} \cdot \varepsilon_h$$

$$W = \sum_{k=1}^{N} X_k = \left(\sum_{k=1}^{N} A^k \right) \cdot X_0 + \sum_{k=1}^{N} \sum_{h=1}^{k} A^{k-h} \cdot \varepsilon_h$$

$$Cov(W) = (1-A)^{-1} \cdot Cov(X_0) \cdot [(1-A)^{-1}]^t + \sum_{k=1}^{N} \Psi_k \cdot \Omega \cdot \Psi_k^t$$

$$\Psi_k = \sum_{h=0}^{N-k} A^k = (1 - A^{N-k+1}) \cdot (1-A)^{-1}$$

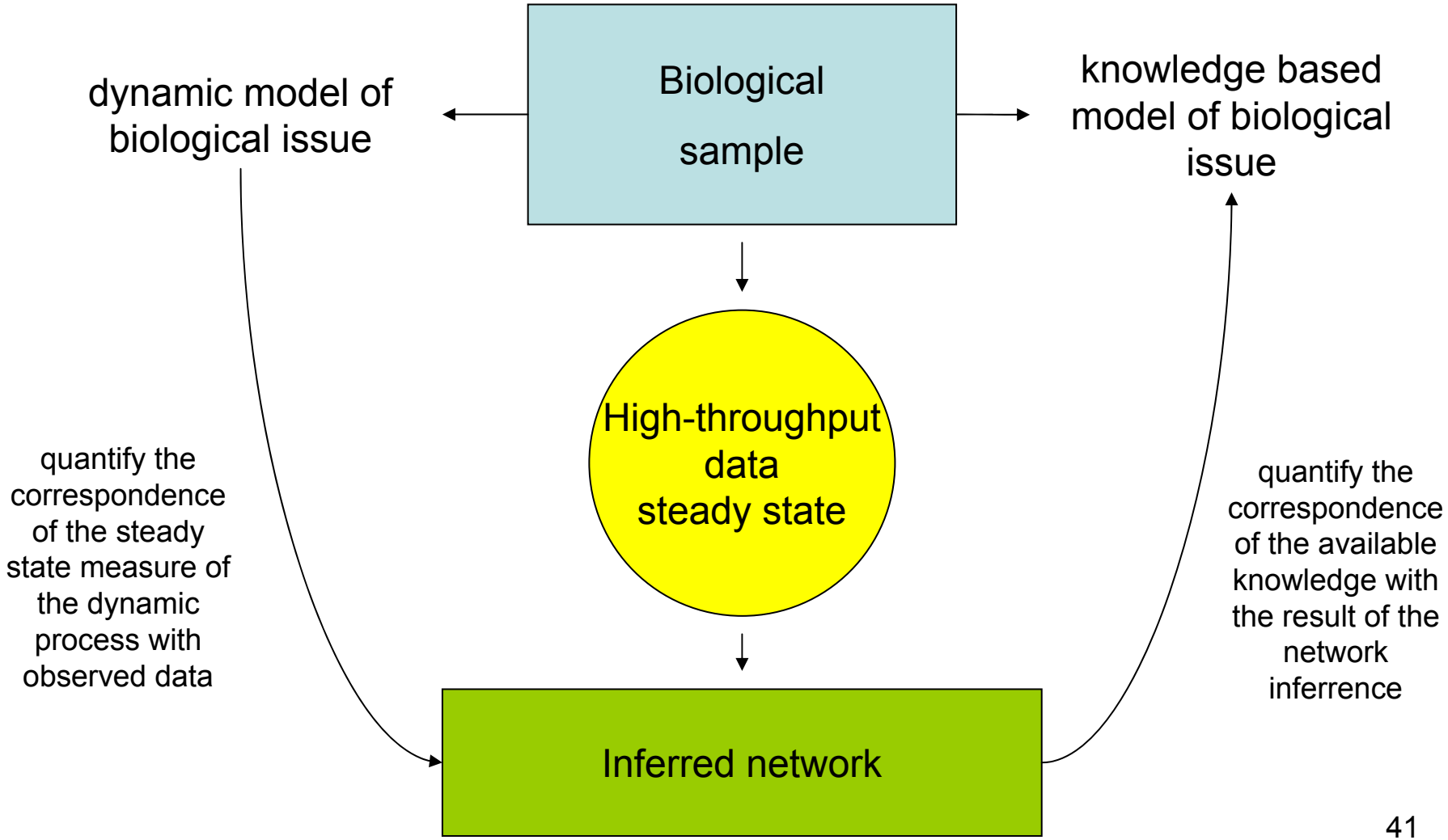
MYC and its translocation: Dynamics

Some inference on the dynamics of MIC-IGH interaction can be made based on specific models.

Unfortunately

1. the literature does not offer detailed analyses on this process, no contribution form systems biology.
2. The model works in the restricted setting (looking exclusively on the small subsystem). As shown before, this approach may suffer from confounding.

Discussion


Biological sample
dynamic model of biological issue
knowledge based model of biological issue
High-throughput data steady state
Inferred network
quantify the correspondence of the steady state measure of the dynamic process with observed data
quantify the correspondence of the available knowledge with the result of the network inferrence

Discussion

- Inferred interaction networks from micorarray gene expression data may not present easily trustworthy biological facts;
- Strategies to validate the biological relevance of these networks are not established yet and even basic strategies are lacking;
- Validation may be approached in two ways: based on formal knowledge derived from the literature or more quantitatively by formal dynamic processes for gene activity in a pathway of interest.
- Biological validation requires closer links of the genomic statistician with bioinformatics and molecular systems biology of the cell.

Discussion

- Notably, methods properly designed to effectively handle the network complexity, dimensionality and heterogeneity, and then deliver interpretable results are still needed.
- Especially biological networks call for further investigation as they are known to suffer from incompleteness and inaccuracy reflected in the high ratios of false positives and false negatives.
- The limitations are in part due to the knowledge of the organisms and in part to the presence of substantial noise floor at both the experimental and the computational (predicted measurements) level.