V12: gene-regulatory networks related to cancerogenesis

What are gene-regulatory networks (GRNs)?

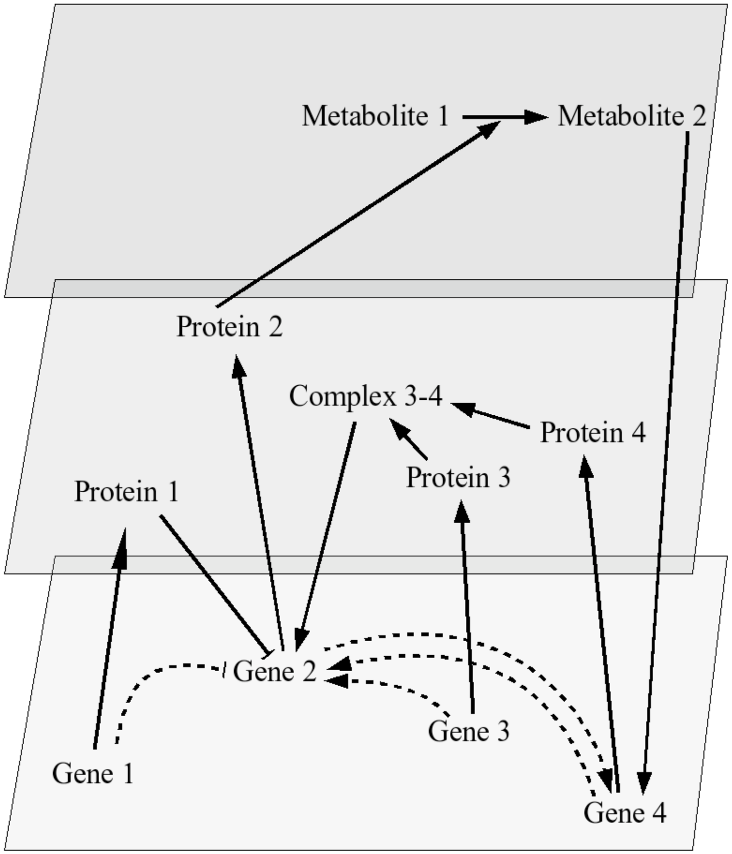
How does one generate GRNs?

Can one set-up GRNs for cancerogenesis based on available data?

What can these GRNs be used for?

Limitations of current GRN models.

Review (bioinformatics III) - GRNs



Example of a gene regulatory network.

Solid arrows: direct associations between genes and proteins (via transcription and translation), between proteins and proteins (via direct physical interactions), between proteins and metabolites (via direct physical interactions or with proteins acting as enzymatic catalysts), and the effect of metabolite binding to genes (via direct interactions).

Lines show direct effects, with arrows standing for activation, and bars for inhibition.

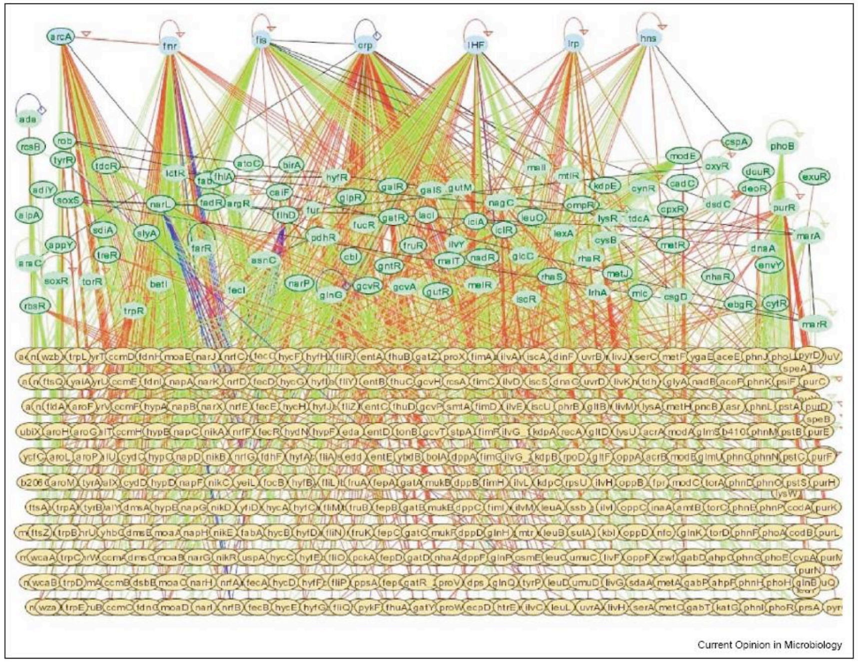
Dashed lines: indirect associations between genes that result from the projection onto 'gene space'. E.g. gene 1 deactivates gene 2 via protein 1 resulting in an indirect interaction between gene 1 and gene 2 (drawn after [Brazhnik00]).

Review (bioinformatics III) – GRN of E. coli

`RegulonDB`: database with information on transcriptional regulation and operon organization in *E.coli*; 105 regulators affecting 749 genes

→ 7 regulatory proteins (CRP, FNR, IHF, FIS, ArcA, NarL and Lrp) are sufficient to directly modulate the expression of more than half of all *E.coli* genes.

→ Out-going connectivity follows a power-law distribution
→ In-coming connectivity follows exponential distribution (Shen-Orr).



Martinez-Antonio, Collado-Vides, Curr Opin Microbiol 6, 482 (2003)

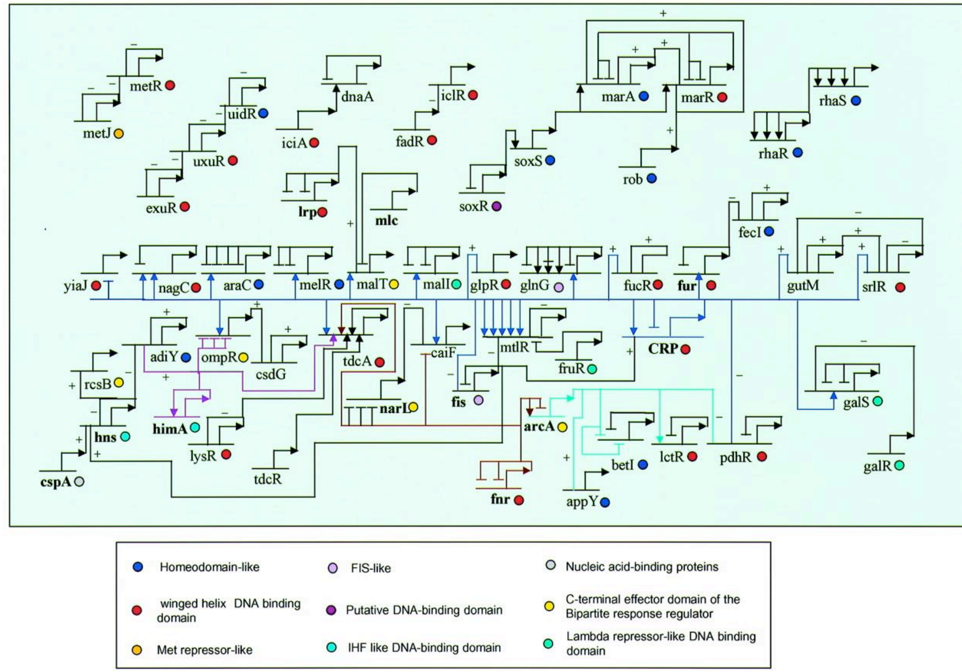
Review (bioinformatics III) – Regulatory cascades in *E.coli*

The TF regulatory network in *E.coli*.
When more than one TF regulates a gene, the order of their binding sites is as given in the figure. An arrowhead is used to indicate positive regulation when the position of the binding site is known.

Horizontal bars indicates negative regulation when the position of the binding site is known. In cases where only the nature of regulation is known, without binding site information, + and – are used to indicate positive and negative regulation.

The DBD families are indicated by circles of different colours as given in the key. The names of global regulators are in bold.

Regulation of transcription factors in E. coli



Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

How does one generate GRNs?

(1.) „by hand“ based on individual experimental observations

(2) Infer GRNs by computational methods from gene expression data

Here we will follow this recent open-access paper

Briefings in Bioinformatics Advance Access published May 21, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 17 doi:10.1093/bib/bbt034

Supervised, semi-supervised and unsupervised inference of gene regulatory networks

Stefan R. Maetschke, Piyush B. Madhamshettiwar, Melissa J. Davis and Mark A. Ragan

Submitted: 19th January 2013; Received (in revised form): 15th April 2013

What is supervision in the context of GRNs?

Unsupervised methods do not use any data to adjust internal parameters.

Supervised methods, on the other hand, exploit all given data to optimize parameters such as weights or thresholds.

Semi-supervised methods use only part of the data for parameter optimization, for instance, a subset of known network interactions.

What is supervision in the context of GRNs?

Inference methods (to infer = *dt. aus etwas ableiten/folgern*) aim to recreate the topology of a genetic regulatory network e.g. based on expression data only.

The **accuracy** of a method is assessed by the extent to which the network it infers is similar to the true regulatory network.

We quantify similarity by the area under the Receiver Operator Characteristic curve (AUC)

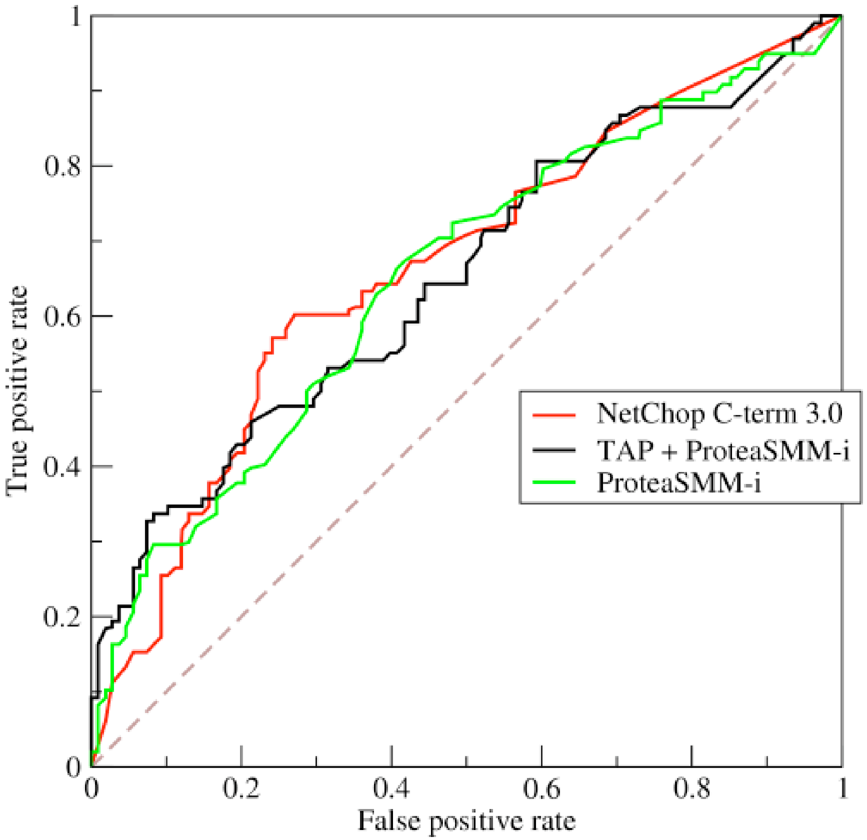
$$AUC = \frac{1}{2}\sum_{k=1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1})$$

where X_k is the false-positive rate and Y_k is the true positive rate for the k-th output in the ranked list of predicted edge weights.

An AUC of 1.0 indicates a perfect prediction, while an AUC of 0.5 indicates a performance no better than random predictions.

AUC

…



Divide data into bins.

Measure value of function Y at midpoint of bin -> factor 0.5

$$AUC = \frac{1}{2}\sum_{k=1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1})$$

www.wikipedia.org

What is supervision in the context of GRNs?

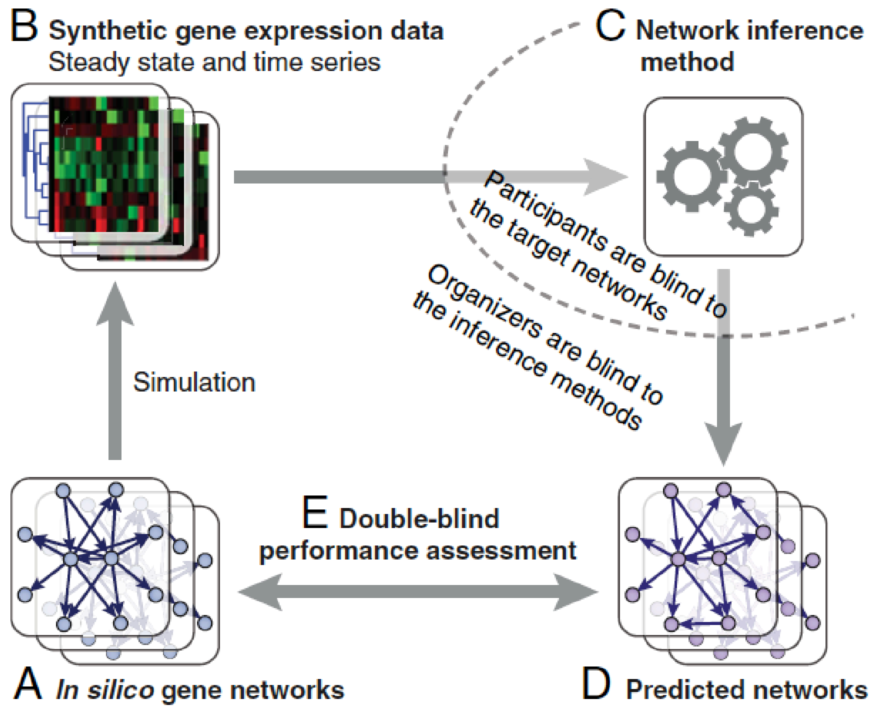
Authors performed evaluations on simulated, steady-state expression data, generated from subnetworks extracted from *E. coli* and *Saccharomyces cerevisiae* networks.

This allowed them to assess the accuracy of an algorithm against a perfectly known true network.

The programs `GeneNetWeaver` and `SynTReN` were used to extract subnetworks and to simulate gene expression data.

Review (bioinfo III):
Mathematical reconstruction of Gene Regulatory Networks

DREAM: **D**ialogue on **R**everse **E**ngineerging **A**ssessment and **M**ethods



Aim:
systematic evaluation of methods for reverse engineering of network topologies (also termed network-inference methods).

Problem:
correct answer is typically not known for real biological networks

Approach:
generate synthetic data

Gustavo Stolovitzky/IBM

Marbach et al. PNAS 107, 6286 (2010)

Review (bioinfo III): Generation of Synthetic Data

Transcriptional regulatory networks are modelled consisting of genes, mRNA, and proteins.

The state of the network is given by the vector of mRNA concentrations x and protein concentrations y.

We model only transcriptional regulation, where regulatory proteins (TFs) control the transcription rate (activation) of genes (no epigenetics, microRNAs etc.).

The gene network is modeled by a system of differential equations

$$\frac{dx_i}{dt} = m_i \cdot f_i(\mathbf{y}) - \lambda_i^{\mathrm{RNA}} \cdot x_i$$

$$\frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{\mathrm{Prot}} \cdot y_i,$$

where m_i is the maximum transcription rate, r_i the translation rate, λ_i^{RNA} and λ_i^{Prot} are the mRNA and protein degradation rates and $f_i(.)$ is the so-called **input function** of gene i.

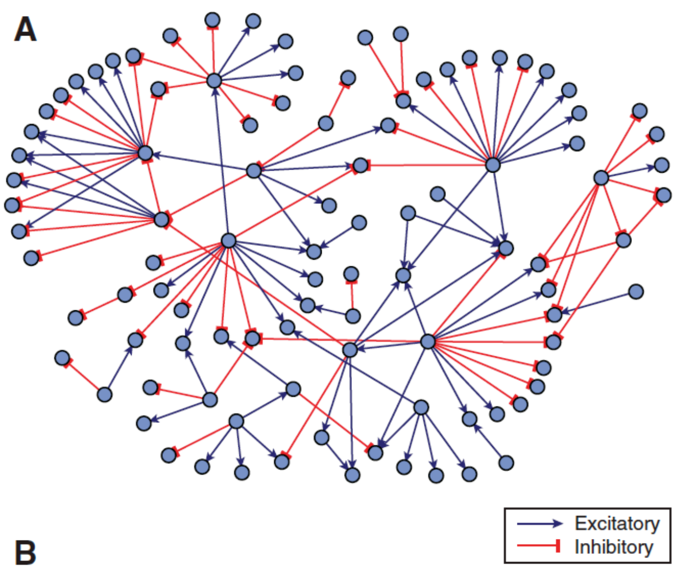
Marbach et al. PNAS 107, 6286 (2010)

Review (bioinfo III): Synthetic networks

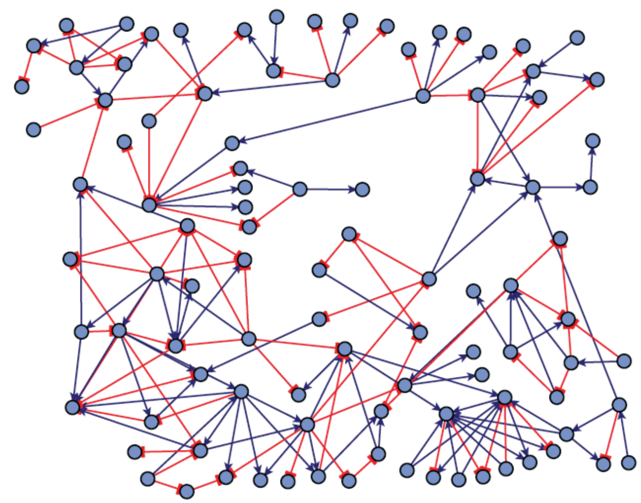
The challenge was structured as three separate subchallenges with networks of 10, 50, and 100 genes, respectively. For each size, five in silico networks were generated.

These resembled realistic network structures by extracting modules from known transcriptional regulatory network for *Escherichia coli* (2x) and for yeast (3x).

Example network *E.coli*



Example network yeast



Marbach et al. PNAS 107, 6286 (2010)

Unsupervised methods

Unsupervised methods are either based on **correlation** or on **mutual information**.
…

Correlation-based network inference methods assume that correlated expression levels between two genes are indicative of a regulatory interaction.

Correlation coefficients range from -1 to 1.

A positive correlation coefficient indicates an activating interaction, while a negative coefficient indicates an inhibitory interaction.

The common correlation measure by Pearson is defined as

$$corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma(X_i) \cdot \sigma(X_j)}$$

where X_i and X_j are the expression levels of genes *i* and *j*, cov(.,.) denotes the covariance, and σ is the standard deviation (see lecture V11) – examples.

Rank-based unsupervised methods

Pearson's correlation measure assumes normally distributed values.
This assumption does not necessarily hold for gene expression data.

Therefore rank-based measures are frequently used.
The measures by Spearman and Kendall are the most common.

Spearman's method is simply Pearson's correlation coefficient for the ranked expression values

Kendall's τ coefficient is computed as

$$\tau(X_i, X_j) = \frac{con(X_i^r, X_j^r) - dis(X_i^r, X_j^r)}{\frac{1}{2}n(n-1)}$$

where X^r_i and X^r_j are the ranked expression profiles of genes *i* and *j*.

Con(.) denotes the number of concordant value pairs (i.e. where the ranks for both elements agree). *dis(.)* is the number of disconcordant value pairs in X^r_i and X^r_j. Both profiles are of length *n*.

WGCNA

WGCNA is a modification of correlation-based inference methods that **amplifies high correlation coefficients** by raising the absolute value to the power of β ('softpower').

$$w_{ij} = |corr(X_i, X_j)|^{\beta}$$

with $\beta \geq 1$.

Because softpower is a nonlinear but monotonic transformation of the correlation coefficient, the prediction accuracy measured by AUC will be no different from that of the underlying correlation method itself.

Unsupervised methods based on mutual information

Relevance networks (RN) introduced by Butte and Kohane measure the **mutual information (MI)** between gene expression profiles to infer interactions.

The MI *I* between discrete variables X_i and X_i is defined as

$$I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log\left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)}\right)$$

where $p(x_i\,, x_j)$ is the **joint probability distribution** of X_i and X_j (both variables fall into given ranges) and
$p(x_i\,)$ and $p(x_i\,)$ are the **marginal probabilities** of the two variables (ignoring the value of the other one).

X_i and X_j are required to be discrete variables.

Unsupervised methods: Z-score

Z-SCORE is a network inference strategy by Prill et al. that takes advantage of knockout data.

It assumes that a knockout affects directly interacting genes more strongly than others.

The z-score z_{ij} describes the effect of a knockout of gene *i* in the *k*-th experiment on gene *j* as the normalized deviation of the expression level X_{jk} of gene *j* for experiment *k* from the average expression $\mu(X_j)$ of gene *j*:

$$z_{ij} = \left| \frac{X_{jk} - \mu(X_j)}{\sigma(X_j)} \right|$$

supervised inference method: SVM

In contrast to unsupervised methods, e.g. correlation methods, the supervised approach does not directly operate on pairs of expression profiles but on feature vectors that can be constructed in various ways.

The authors computed the outer product of two gene expression profiles X_i and X_j to construct feature vectors:

$$\mathbf{x} = X_i X_j^T$$

A sample set for the training of the SVM is then composed of feature vectors $\mathbf{x}_i$ that are labeled γ_i = +1 for gene pairs that interact and γ_i = -1 for those that do not interact.

Labelling of samples

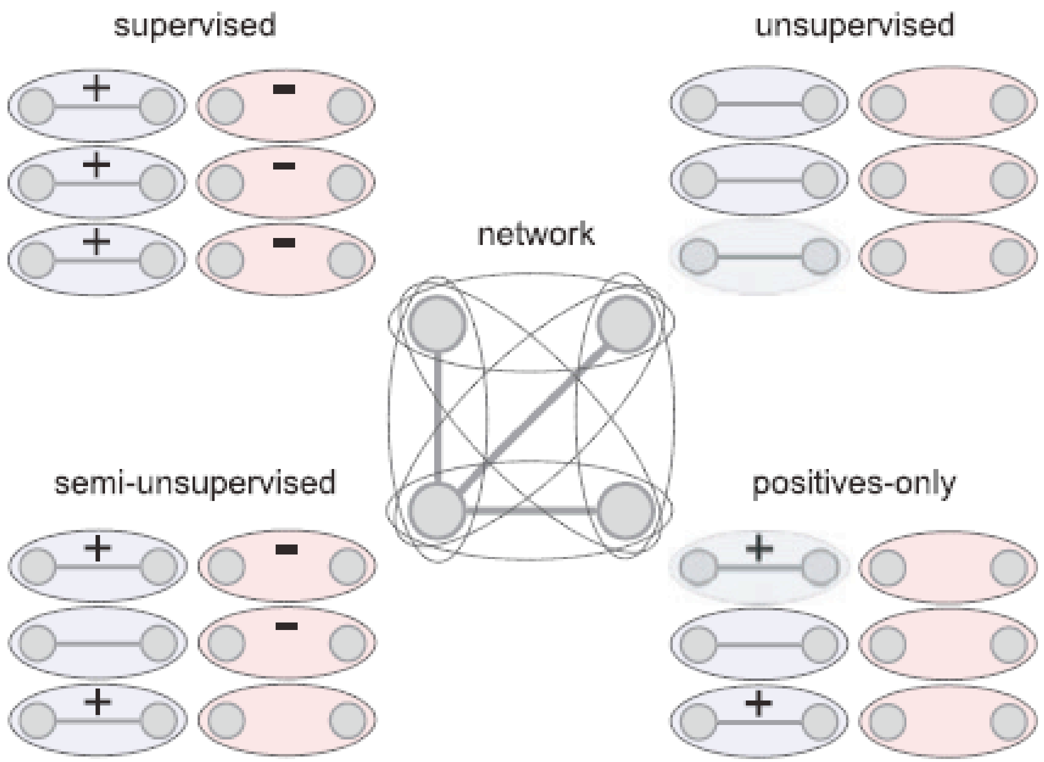
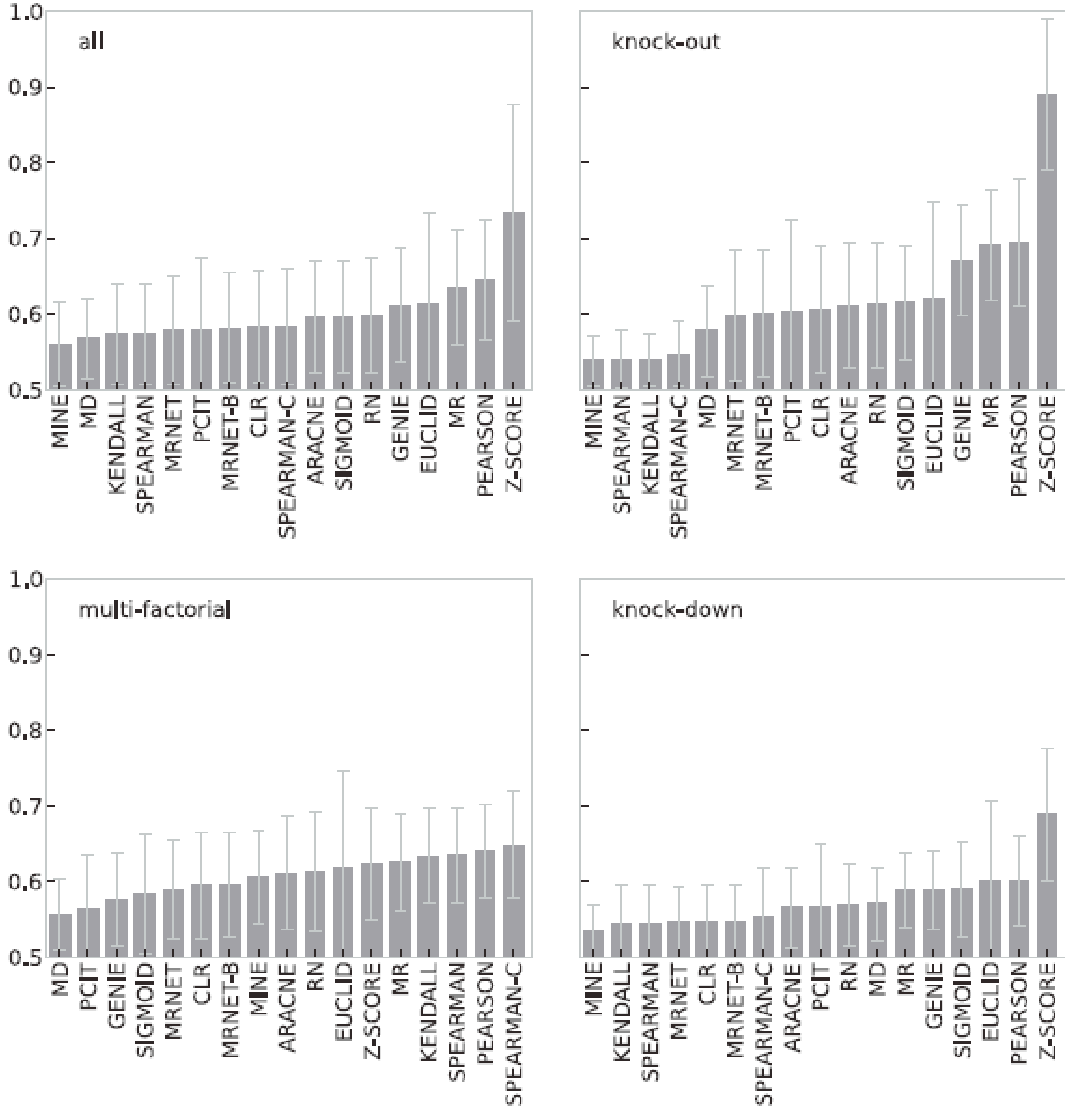


Figure 2: Original labeling of samples for supervised, unsupervised, semi-supervised and positives-only prediction methods. All the six samples within a sample set are generated by a four-node network with three interactions.

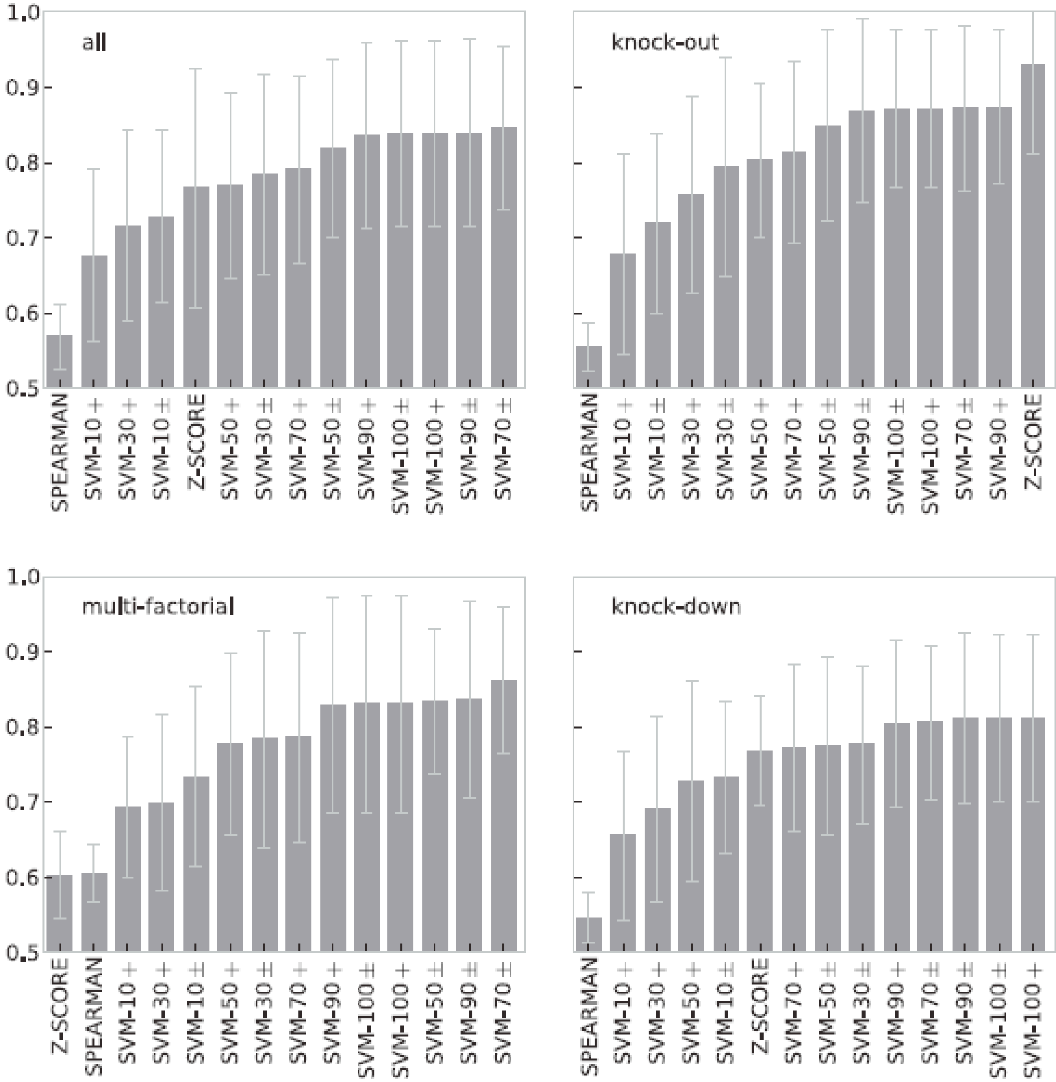
Obtained AUC curves for unsupervised methods



A simple Pearson‘s correlation gives the second-best performance.

Except for the *z*-score method, accuracies are generally low.

Comparison with supervised method



Supervised learning methods give much better results than unsupervised methods.

(10, 30, … 100 indicates the percentage of labeled data).

The only exception is the excellent performance of the z-score method for knock-out data.

Application to ovarian cancer data

Madhamshettiwar *et al. Genome Medicine* 2012, 4:41
http://genomemedicine.com/content/4/5/41

Genome Medicine

RESEARCH Open Access

Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets

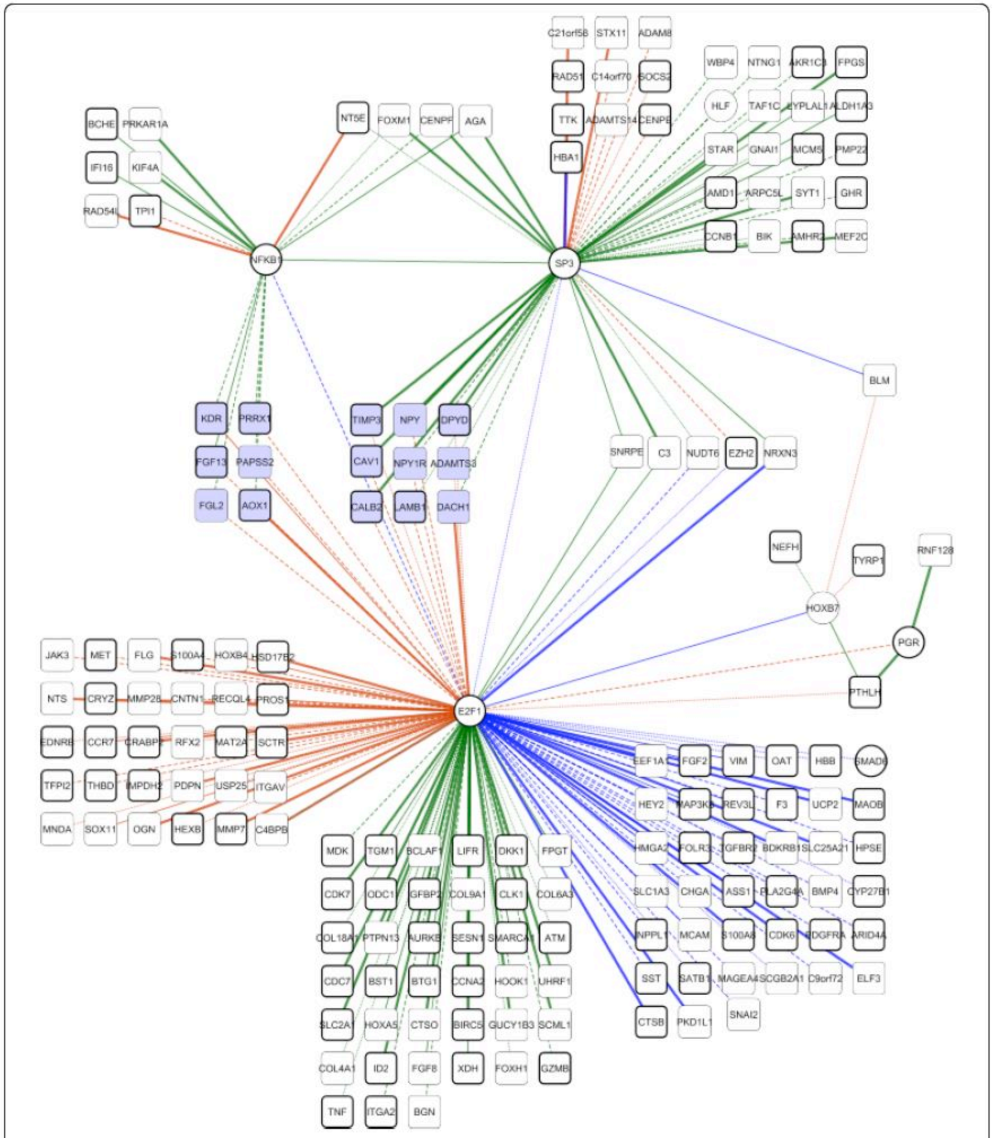
Piyush B Madhamshettiwar[1,2], Stefan R Maetschke[1,2], Melissa J Davis[1,2,3], Antonio Reverter[4] and Mark A Ragan[1,2*]

Application to ovarian cancer data

Table 1 Accuracies of unsupervised and supervised GRNI methods on different datasets

	Unsupervised method		SIRENE
Datasets	**Method**	**AUC**	**AUC**
DREAM3 (knockdown): genes 100, samples 100	MRNET	0.59	0.71
DREAM4 (multifactorial): genes 100, samples 100	GENIE	0.79	0.69
Ovary normal: genes 2,450, samples 12	RN	0.55	0.62
Ovary normal: genes 282, samples 12	RN	0.70	0.86

Application to ovarian cancer data



The ovarian gene regulatory network inferred using the program SIRENE, showing target genes (rectangles) and TFs (circles). 2 clusters of genes (shaded blue, in the centre of the figure) switch regulators between the two conditions, controlled by SP3 or NFB1 in normal and by E2F1 in cancer.

Bold nodes are known to have protein products that are targeted by anti-cancer drugs.

Edge colors: green, normal; orange, cancer; blue, both.

Edge line type: bold, literature and TFBS; solid, literature; dashed, TFBS; dotted, no evidence.

Application to ovarian cancer data

To identify the proteins regarded as anti-cancer drug targets, we input all 178 proteins from our GRN to CancerResource.

61% of the proteins from our network are targeted by at least one anticancer drug.

In many cases a single drug targets multiple proteins, or conversely multiple drugs target a single protein.

Application to ovarian cancer data

Table 2 Druggability analysis results

Gene name	Gene type	Targeted drugs
Top 10 target genes		
BCHE	Enzyme	Bicalutamide, genistein, choline, isoflurophate, hexafluorenium, demecarium bromide, echothiophate iodide, butyric acid
CDK7	Protein kinase	Lycopene, genistein, flavopiridol
DKK1	Receptor ligand	Decitabine
CCR7	GPCR	Decitabine
TPI1	Enzyme	Fluorouracil, quercetin
HSD17B2	Enzyme	NADH
HBB	Transporter	Iron dextran complex
Angiogenesis genes: SP3 targets		
TIMP3	Binding protein	Salinomycin, decitabine, sulindac, adaphostin
CAV1	Binding protein	Decitabine, progesterone, mifepristone
CALB2	Binding protein	Oxaliplatin, fluorouracil
LAMB1	Receptor ligand	Benzamidine, carebastine, anistreplase, tenecteplase
DPYD	Enzyme	Oxaliplatin, gemcitabine, docetaxel, s1(combination), capecitabine, cisplatin, fluorouracil, tegafur, carboplatin, paclitaxel, genistein, enfuvirtide, raltitrexed, amifostine, irinotecan, methotrexate, mitoguazone, uracil
Angiogenesis genes: NFκB1 targets		
KDR	Receptor with enzyme activity	Epigallocatechin gallate, resveratrol, sorafenib, sunitinib, bevacizumab, sirolimus, conivaptan, zonampanel, SU6668, vatalanib, vandetanib, axitinib, cediranib, trapoxin, motesanib, E 7080, erlotinib, Ca0456456, geldanamycin
FGF13	Receptor ligand	Bicalutamide
PRRX1	Transcription factor	Alitretinoin
AOX1	Enzyme	Isovanillin, norcantharidin, NSC336628

Genes and anti-cancer drugs targeting their products were obtained using Cancer Resource and PharmGKB webtools and databases. GPCR, G-protein-coupled receptor.

Summary

Network inference is a very important active research field.

Inference methods allow to construct the topologies of gene-regulatory networks solely from expression data (unsupervised methods).

Supervised methods show far better performance.

Performance on real data is lower than on synthetic data because regulation in cells is not only due to interaction of TFs with genes, but also depends on epigenetic effects (DNA methylation, chromatin structure/histone modifications, and miRNAs).