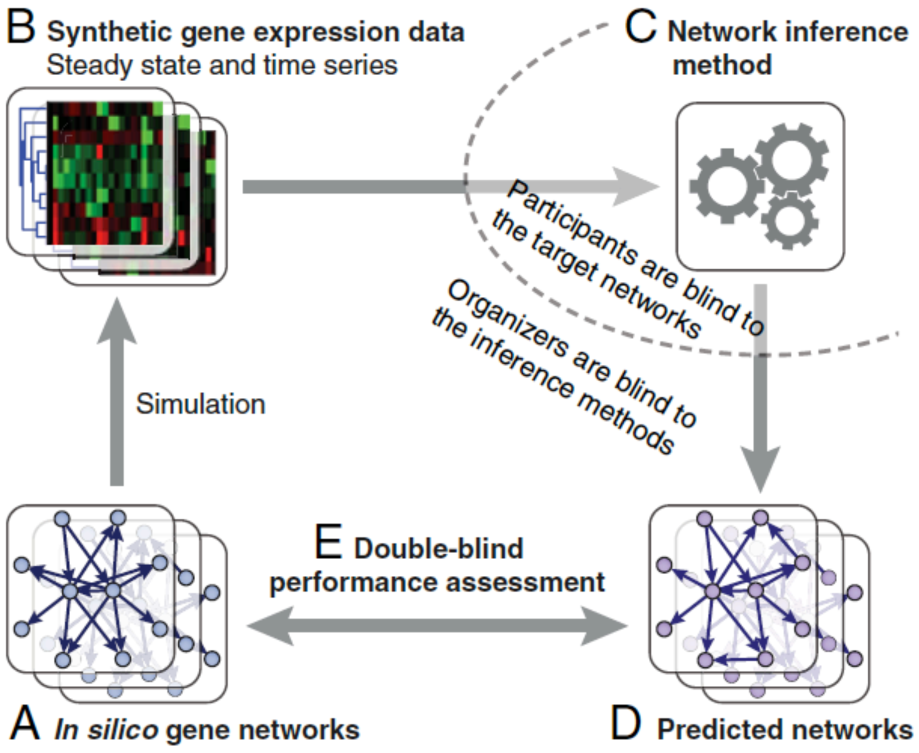
Mathematical reconstruction of Gene Regulatory Networks

DREAM: Dialogue on Reverse Engineering Assessment and Methods

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

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

Approach:
generate **synthetic data**


B Synthetic gene expression data
Steady state and time series
C Network inference method
Participants are blind to the target networks
Organizers are blind to the inference methods
Simulation
E Double-blind performance assessment
A *In silico* gene networks
D Predicted networks

Gustavo Stolovitzky/IBM

Marbach et al. PNAS 107, 6286 (2010)

Generation of Synthetic Data

Model transcriptional regulatory networks consisting of mRNA and proteins.

Current **state** of network :

vector of **mRNA concentrations x** and **protein concentrations y**.

Considered is only transcriptional regulation, where regulatory proteins (TFs) control the activation of genes; no epigenetics, microRNAs etc.

The gene network is modeled by a **system of differential equations** (equivalent to V11, slide 24).

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

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

m_i : maximum **transcription rate**,

r_i : **translation rate**,

$f_i(.)$: so-called **input function** of gene *i*.

λ_i^{RNA} , λ_i^{Prot} : mRNA and protein **degradation rates**

Marbach et al. PNAS 107, 6286 (2010)

The input function $f_i()$

The input function describes the relative activation of a gene given the transcription-factor (TF) concentrations **y**.
Its value is between 0 (gene shut off) and 1 (gene maximally activated).

We assume that **binding of TFs** to cis-regulatory sites on the DNA is in **quasi-equilibrium**, since TF binding is orders of magnitudes faster than transcription and translation (which take minutes).

In the **simplest case**, a gene *i* is regulated by a single TF *j*.

In this case, its promoter has only two states:
either the TF is bound (state *S1*) or not bound (state *S0*).

The probability $P(S_1)$ that the gene *i* is in state *S1* at a particular moment is given by the **fractional saturation**, which depends on the TF concentration y_j

Marbach et al. PNAS 107, 6286 (2010)

Excursion: the Hill equation (see V9, slide 33)

Let us consider the binding reaction of two molecules *L* and *M*:

$$L + M \leftrightarrows LM$$

The **dissociation equilibrium constant** K_D is defined as:

$$K_\mathrm{D} = \frac{[L][M]}{[LM]}$$

where [*L*], [*M*], and [*LM*] are the molecular concentrations of *L* and *M* and of the complex *LM*.

In equilibrium, we may take *T* as the total concentration of molecule *L*

$$T = [L] + [LM]$$

y is the **fraction** of molecules *L* **that have reacted (bound)**

$$y = \frac{[LM]}{[LM] + [L]}$$

Goutelle et al. Fundamental & Clinical Pharmacology 22 (2008) 633–648

Excursion: the Hill equation (see V9, slide 34)

$$y = \frac{[LM]}{[LM] + [L]}$$

Substituting $[LM]$ by $[L]\,[M] / K_D$ gives (rearranged from $K_D = \frac{[L][M]}{[LM]}$)

$$y = \frac{([L]\,[M]\,/K_D)}{([L]\,[M]\,/K_D) + [L]} = \frac{([M]\,/K_D)}{([M]\,/K_D) + 1}$$

Back to our case about TF binding to DNA. **(slightly different from V9)**

TF *j* then takes the role of *M*. Its concentration is y_j.

The probability $P(S_1)$ that the gene *i* is in state *S1* at a particular moment is given by the *fractional saturation*, which depends on the TF concentration y_j

$$P\{S_1\} = \frac{\chi_j}{1 + \chi_j} \quad \text{with} \quad \chi_j = \left(\frac{y_j}{k_{ij}}\right)^{n_{ij}}$$

k_{ij} : dissociation constant for TF *j* at the promoter of gene *i*

n_{ij} : Hill coefficient (describing cooperativity) for this binding equilibrium.

The input function $f_i()$

$$P\{S_1\} = \frac{\chi_j}{1+\chi_j} \quad \text{with} \quad \chi_j = \left(\frac{y_j}{k_{ij}}\right)^{n_{ij}}$$

$P(S_1)$ is large if the concentration y_j of TF j is large
and if the dissociation constant k_{ij} is small (strong binding).

The bound TF either activates or represses the expression of the gene.

In state S_0 the **relative activation** is α_0. In state S_1 it is α_1.

The **input function** $f_i(y_j)$ is obtained from $P(S_1)$ and its complement $P(S_0)$.

$$P\left(S_0\right) = 1 - \frac{\chi_j}{1+\chi_j} = \frac{1+\chi_j-\chi_j}{1+\chi_j} = \frac{1}{1+\chi_j}$$

The input function describes the **mean activation** of gene i as a function of the TF concentration y_j

$$f(y_j) = \alpha_0 P\{S_0\} + \alpha_1 P\{S_1\} = \frac{\alpha_0 + \alpha_1\chi_j}{1+\chi_j}$$

Marbach et al. PNAS 107, 6286 (2010)

The input function $f_i()$

This approach can be generalized
to an **arbitrary number** of regulatory inputs.

A gene that is controlled by *N* TFs has 2^N states:
each of the TFs can be bound or not bound.

Thus, the input function for *N* regulators is

$$f(\mathbf{y}) = \sum_{m=0}^{2^N-1} \alpha_m P\{S_m\}$$

Marbach et al. PNAS 107, 6286 (2010)

Correlation-based unsupervised methods

Correlation-based network inference methods assume that correlated expression levels between two genes are indicative of a regulatory interaction (note however slide 42 in lecture V9).

Correlation coefficients range from -1 to 1.
A **positive** correlation coefficient indicates an **activating interaction**, whereas 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**.

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 :

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

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.

Z-score

Z-SCORE is a network inference strategy by Prill *et al.* that assumes the availability of **knockout experiments** that lead to a change in other genes.

The assumption is that the knocked-out gene *i* in experiment *k* affects more strongly the genes that it regulates than the others.

The effect of gene *i* on gene *j* is captured with the Z-score z_{ij}:

$$z_{ij} = \left| \frac{x_{jk} - \mu_{X_j}}{\sigma_{X_j}} \right|$$

assuming that the *k*-th experiment is a knockout of gene *i*, μ_{Xj} and σ_{Xj} are respectively the mean and standard deviation of the empirical distribution of the expression values x_{jk} of gene *j*.

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 between discrete variables (here: genes) X_i and X_j is defined as

$$M_{ij} = \sum_{X_i} \sum_{X_j} p(X_i, X_j) \log_2 \frac{p(X_i, X_j)}{p(X_i)p(X_j)}$$

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 $p(X_j)$ are the **marginal probabilities** of the two variables (ignoring the value of the other one).

RELNET

The RELNET is the simplest method based on **mutual information**.

For each pair of genes, the mutual information M_{ij} is estimated and
the edge between genes *i* and *j* is created
if the mutual information is above a threshold.

Although mutual information is more general than Pearson correlation,
in practice both give similar results.

Bellot *et al. BMC Bioinformatics* (2015) 16:312

CLR

The Context Likelihood or Relatedness network (CLR) method is an extension of RELNET.

CLR derives a score that is associated to the empirical distribution of the mutual information values.

The score between gene *i* and gene *j* is:

$$c_{ij} = \sqrt{c_i^2 + c_j^2}, \text{ with } c_i = \max\left(0, \frac{M_{ij} - \mu_{M_i}}{\sigma_{M_i}}\right) \text{ and}$$

$$c_j = \max\left(0, \frac{M_{ji} - \mu_{M_j}}{\sigma_{M_j}}\right).$$

with the mean μ_{Mi} and standard deviation σ_{Mi} of the empirical distribution of the mutual information between these genes and other genes,

$$\mu_{M_i} = \frac{1}{G}\sum_{l=1}^{G} M_{il}, \quad \sigma_{M_i} = \sqrt{\frac{1}{G-1}\sum_{l=1}^{G}(M_{il} - \mu_{M_i})^2}$$

Bellot *et al. BMC Bioinformatics* (2015) 16:312