Inference of Gene Regulatory Networks: Inference, Analysis and Interpretation

Frank Emmert-Streib

Computational Biology and Machine Learning Lab
Center for Cancer Research and Cell Biology
Queen’s University Belfast, UK

3th August 2012, COST Summer School in Bioinformatics 2012
Availability of the Tutorial

• Web: www.bio-complexity.com
• Email: v@bio-complexity.com
Part I

Motivation
Why infer networks?
Central Dogma of Molecular Biology
[17]
<table>
<thead>
<tr>
<th>gene network</th>
<th>node</th>
<th>edge</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcription regulation net.</td>
<td>protein/gene</td>
<td>protein-DNA</td>
<td>[5, 7, 29]</td>
</tr>
<tr>
<td>protein network</td>
<td>protein</td>
<td>PPI</td>
<td>[23, 44, 45]</td>
</tr>
<tr>
<td>signaling network</td>
<td>protein/gene</td>
<td>PPI</td>
<td>[12, 28, 37]</td>
</tr>
<tr>
<td>metabolic networks</td>
<td>metabolite</td>
<td>reaction</td>
<td>[21, 27, 31]</td>
</tr>
<tr>
<td>regulatory network</td>
<td>protein/gene</td>
<td>multiple</td>
<td>[34, 38, 40]</td>
</tr>
</tbody>
</table>

**Table:** Different types of gene networks and their characteristics.
Table: Different types of gene networks and their characteristics.

<table>
<thead>
<tr>
<th>gene network</th>
<th>node</th>
<th>edge</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcription regulation net.</td>
<td>protein/gene</td>
<td>protein-DNA</td>
<td>[5, 7, 29]</td>
</tr>
<tr>
<td>protein network</td>
<td>protein</td>
<td>PPI</td>
<td>[23, 44, 45]</td>
</tr>
<tr>
<td>signaling network</td>
<td>protein/gene</td>
<td>PPI</td>
<td>[12, 28, 37]</td>
</tr>
<tr>
<td>metabolic networks</td>
<td>metabolite</td>
<td>reaction</td>
<td>[21, 27, 31]</td>
</tr>
<tr>
<td>regulatory network</td>
<td>protein/gene</td>
<td>multiple</td>
<td>[34, 38, 40]</td>
</tr>
</tbody>
</table>

Regulatory networks are inferred from gene expression data.

Two different types of networks:

- **inferential networks**: regulatory networks
- **phenomenological networks**: protein network, metabolic network etc
Motivation

- DNA
- genotype
- phenotype
- interaction
- environment
Motivation

phenotype

gene networks

genotype

DNA

interaction

environment
Motivation

Gene networks - bottleneck

phenotype

interaction

DNA

environment

genotype
Summary of poster submissions to ISMB 2011

Network Biology 23.7%
Databases 11.4%
Functional Genomics 7.8%
Sequence analysis 16.1%
Structural Genomics 1.7%
Proteomics 3.4%
Microarrays 9.5%
Evolution 5.3%
Text Mining 3.4%
Ontologies 2.9%
Population Genetics and Variation 2.6%
Machine Learning 5.8%
Comparative Genomics 5.8%
How to infer networks?
Practical approach (tools of the trade)

Statistical programming language R

1. freely available (http://www.r-project.org/) for Linux, Mac, Windows
2. similar to S (developed in the 80s in the Bell Labs)
3. gold standard in Biostatistics/Computational Biology
4. flexible yet simple (compared to perl or C/C++)
5. many packages are available:
   - CRAN: http://cran.r-project.org/web/packages/
   - Bioconductor: http://www.bioconductor.org/
Overall outline of the tutorial

1. Motivation
2. Interpretation of regulatory networks
3. Network Inference Methods
4. Evaluation measures
5. Data
   - biological data (DNA microarray, RNA-seq)
   - simulated data
6. Practical examples - using R (statistical programming language)
7. Applications
   - comparison of different inference methods
   - B-cell lymphoma
8. Summary
Part II

Interpretation of regulatory networks
graph
(undirected, unweighted, labeled)
gene network
Interpretation of regulatory networks

gene expression data → regulatory network

interaction → protein-DNA binding
protein-protein interaction → gene

gene product
Interpretation of regulatory networks

Gene expression data → regulatory network

interaction

protein-DNA binding
protein-protein interaction

gene

gene product
Are both networks the same?

- From a mathematical point of view:
  Both networks contain the same amount of information.

- From a biological point of view:
  The right figure contains more biological information.
Are both networks the same?

Left network is an **abstraction** of reality $\implies$ Model (based on definitions)
Despite the fact that gene regulatory network interaction protein-DNA binding protein-protein interaction gene product gene expression data regulatory network has an intuitive **biological interpretation** it is also a **model** (mathematical representation).

In general, gene networks have a dual meaning:
- interpretation
- model
Despite the fact that gene regulatory network interaction protein-DNA binding protein-protein interaction gene product gene expression data regulatory network

has an intuitive **biological interpretation** it is also a **model** (mathematical representation).

In general, gene networks have a dual meaning:

- **interpretation**
- **model** → **networks are inherently abstract**
Part III

Network Inference Methods
Overview of part III

- general overview of network inference methods [18]
- distinction between association and causal networks
- discussion of the principle working mechanism of some network inference methods
## Network Inference Methods

<table>
<thead>
<tr>
<th>Correlation-based</th>
<th>Mutual Information-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-expression networks</td>
<td>RN (relevance networks)</td>
</tr>
<tr>
<td>Low-order partial correlation</td>
<td>C3NET (conservative causal core)</td>
</tr>
<tr>
<td>GGM (graphical Gaussian model)</td>
<td>ARACNE</td>
</tr>
<tr>
<td></td>
<td>CLR (context likelihood of relatedness)</td>
</tr>
<tr>
<td></td>
<td>MRNET (maximum relevance, minimum redundancy)</td>
</tr>
</tbody>
</table>

Frank Emmert-Streib

Gene Regulatory Networks
Network Inference Methods

Co-expression networks
Low-order partial correlation
GGM (graphical Gaussian model)
correlation-based

RN (relevance networks)
ARACNE
CLR (context likelihood of relatedness)
C3NET (conservative causal core)
MRNET (maximum relevance, minimum redundancy)
mutual information-based

association networks  causal networks
Causal (gene) network:

- an ‘edge’ in the network corresponds to a *biochemical interaction* that can be *validated experimentally*

Examples for causal interactions:
Causal (gene) network:

- an ‘edge’ in the network corresponds to a biochemical interaction that can be validated experimentally

Example for an association:
Defining property

Causal (gene) network:
- an ‘edge’ in the network corresponds to a biochemical interaction that can be validated experimentally

Problems:
- conducting a wet-lab experiments for two genes that do not directly interact with each other requires guesswork
- example: differentially expression of genes
Defining property

Causal (gene) network:
- an ‘edge’ in the network corresponds to a biochemical interaction that can be validated experimentally for example: protein-DNA binding, protein-protein interaction

Potential application: study individual networks
Potential application: comparison of gene networks ⇒ identify (causal) interaction changes (between conditions; disease stages or grades etc)

-do the networks look similar
Inference methods
Co-expression networks

The correlation $r_{ij}$ (between gene $i$ and $j$) is filtered according to:

The sigmoid function,

$$A_{ij} = \frac{1}{1 + \exp(-\alpha (r_{ij} - \tau_0))},$$

and the power adjacency function,

$$A_{ij} = |r_{ij}|^\beta.$$

Both types of adjacency functions lead to undirected but weighted networks.

It has been suggested to choose the above parameters in a way that the resulting network has approximately a scale-free degree distribution [50].
Low-order partial correlation

Starting from a fully connected adjacency matrix $A$ [14].

**Zero-order:**
- If $\rho_{ij} = \rho(X_i, X_j) = 0$ (statistical test) \(\implies\) the connection between gene $X_i$ and $X_j$ is deleted, $A_{ij} = A_{ji} = 0$.

**First-order:**
- $\rho_{ij|k} = \rho_{X_i,X_j|X_k}$, are calculated for all triplets of genes with $\rho_{ij} \neq 0$.
- If there is at least one gene $X_k$ for which $\rho_{ij|k} = 0$ holds (hypothesis test) \(\implies\) the connection between gene $X_i$ and $X_j$ is deleted, $A_{ij} = A_{ji} = 0$.

**n-th-order:**
- analogously

The resulting network from this procedure is frequently called an **undirected dependency graph (UDG)** [42].
Assumption: best-model configuration

\[
A_0 = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix}
\]

initialization

\[
A_1 = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix}
\]

correlation of zero-order

fully connected

true network

estimated network \((A^1)\)
Assumption: best-model configuration

\[ A_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \]

correlation of zero-order

\[ A_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \]

correlation of first-order

true network

estimated network \((A_2)\)
The partial correlation of first order:

\[ r_{x_i x_j | x_k} = \frac{r_{x_i x_j} - r_{x_i x_k} r_{x_j x_k}}{\sqrt{1 - r_{x_i x_k}^2} \sqrt{1 - r_{x_j x_k}^2}} \]
A GGM is given by a specific structure of the inverse of the covariance matrix, \( \Omega = \Sigma^{-1} \) (precision or concentration matrix). Network inference methods based on GGM utilize the relation,

\[
\rho_{ij \mid V \setminus \{ij\}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii} \omega_{jj}}},
\]

connecting the partial correlation of full-order with the elements of \( \Omega \), \( \omega_{ij} \in \Omega \).

Start with an unconnected adjacency matrix \( A \).

- If \( \rho_{ij \mid V \setminus \{ij\}} \neq 0 \) (hypothesis test) we include an edge, \( A_{ij} = A_{ji} = 1 \), otherwise there is no edge between \( i \) and \( j \).
The principle idea of RN [8] is to compute all mutual information (MI) values for all pairs of genes and declare mutual information values as relevant if their corresponding value is larger than a given threshold $I_0$. Start with a fully connected adjacency matrix $A$.

- If $I_{ij} = 0$ (hypothesis test) $\implies$ the connection between gene $X_i$ and $X_j$ is deleted, $A_{ij} = A_{ji} = 0$. 

RN (relevance networks)
Mutual information:

\[ I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]
ARACNE (algorithm for the reconstruction of accurate cellular networks)

ARACNE [6, 33]
Start with a fully connected adjacency matrix $A$:

- If $I_{ij} = 0$ (hypothesis test) $\implies$ the connection between gene $X_i$ and $X_j$ is deleted, $A_{ij} = A_{ji} = 0$.

- DPI (data processing inequality): testing all gene-triplets (three genes with significant MI values) such that, for each triplet $(ijk)$, the edge corresponding to the lowest mutual information value $I_1 = I_{i'j'}$, with $(i'j') = \arg\min\{I_{ij}, I_{jk}, I_{ik}\}$, is eliminated from the adjacency matrix, if it is lower than the second lowest MI value $I_2$ multiplied by a factor.

$$A_{i'j'} = A_{j'i'} = \begin{cases} 0 & I_{i'j'} \leq I_2 (1 - \epsilon) \\ \text{unchanged} & \text{otherwise} \end{cases}$$
Network Inference Methods

DPI with epsilon

delete edge with certain probability
(controlled by epsilon)

Frank Emmert-Streib
Gene Regulatory Networks
CLR (context likelihood of relatedness)

For each gene pair \((ij)\) two z-scores are obtained, one for gene \(i\) and one for gene \(j\), by comparing the mutual information value \(I_{ij}\) with gene specific distributions, \(p_i\) and \(p_j\). Here the two distributions \(p_i\) and \(p_j\) correspond to the distributions of mutual information values related to gene \(i\) (\(\{I_{ik}|k \in V\}\)) and gene \(j\) (\(\{I_{jk}|k \in V\}\)). By making a normality assumption about these distributions, corresponding z-scores, \(z_i\) and \(z_j\), can be obtained from which the joint likelihood measure

\[
Z_{ij} = \sqrt{z_i^2 + z_j^2}
\]

is calculated [19].
From [19].
C3NET (conservative causal core)

**C3NET** consists of two main steps [1].

1. elimination of nonsignificant edges (testing for MI=0)
2. select for each gene the edge among the remaining ones with maximum mutual information value

The first step is similar to RN [8], ARACNE [33] or CLR [19] essential for eliminating nonsignificant links, according to a chosen significance level $\alpha$, between gene pairs. In the second step, the most significant link for each gene is selected. This link corresponds also to the highest MI value among the neighbor edges for each gene. This implies that the highest possible number of edges that can be inferred by C3NET is equal to the number of genes under consideration.
principle of C3NET
Software:

- **R package:** c3net
- Implementation of C3NET
- Available from the CRAN repository.


MRNET (maximum relevance, minimum redundancy)

MRNET [36] is an iterative algorithm that identifies potential interaction partners of a target gene \( Y \) that maximize a scoring function.

\[
X_j^s = \arg\max_{X_j \in V \setminus S} (s_j) 
\]

\[
s_j = I(X_j; Y) - \frac{1}{|S|} \sum_{X_k \in S} I(X_j; X_k). 
\]

When a gene, \( X_j \), is found with a score that maximizes Eqn. 1 and \( s_j \) is above a threshold, \( s_0 \), then this gene is added to the set \( S \).

Basic idea: **maximal relevance** (first term in Eqn. 2) for \( Y \), but **minimum redundancy** (second term in Eqn. 2) with respect to the already found interaction partners in the set \( S \). Starting with a fully connected network MRNET reduces successively edges between \( Y \) and \( V \setminus S \) that have not been found by the algorithm.
\[ s_j = I(X_j; Y) - (1 \backslash S) \sum_{X_k \text{in } S} I(X_j; X_k) \]

- principle of MRNET
Software:

- R package: minet
- Implementation of MRNET, ARACNE, CLR
- Available from Bioconductor.

How about Bayesian Networks?

How about Bayesian Networks?


Very high computational complexity (NP-complete) [11].
## Network Inference Methods

<table>
<thead>
<tr>
<th>name/publication</th>
<th>key method</th>
<th>software</th>
</tr>
</thead>
<tbody>
<tr>
<td>coexpression [9, 50]</td>
<td>correlation</td>
<td>WGCNA</td>
</tr>
<tr>
<td>Asymmetric-N [10]</td>
<td>correlation, ranking</td>
<td>no</td>
</tr>
<tr>
<td>SEM [49]</td>
<td>+ , genetic algorithm</td>
<td>no</td>
</tr>
<tr>
<td>ParCorA [14]</td>
<td>partial correlation first order</td>
<td>ParCorA</td>
</tr>
<tr>
<td>GGM [41]</td>
<td>full order partial correlation</td>
<td>GeneNet</td>
</tr>
<tr>
<td>BN [24]</td>
<td>partial correlation n-th order</td>
<td>yes(^1)</td>
</tr>
<tr>
<td>RN [8]</td>
<td>mutual information</td>
<td>MINET</td>
</tr>
<tr>
<td>ARACNE [33]</td>
<td>mutual information, DPI</td>
<td>MINET</td>
</tr>
<tr>
<td>CLR [19]</td>
<td>mutual information, background</td>
<td>MINET</td>
</tr>
<tr>
<td>C3NET [1, 3]</td>
<td>maximum mutual information</td>
<td>c3net</td>
</tr>
<tr>
<td>SA-CLR [47]</td>
<td>mutual information, synergy</td>
<td>no</td>
</tr>
<tr>
<td>MRNET [36]</td>
<td>mutual informations</td>
<td>MINET</td>
</tr>
<tr>
<td>MI3 [47]</td>
<td>three-way mutual information</td>
<td>MI3</td>
</tr>
<tr>
<td>CMI [43]</td>
<td>conditional mutual information</td>
<td>no</td>
</tr>
<tr>
<td>MI-CMI [30]</td>
<td>MI, conditional mutual information</td>
<td>no</td>
</tr>
<tr>
<td>[51]</td>
<td>MI, conditional mutual information</td>
<td>no</td>
</tr>
<tr>
<td>BC3NET [15]</td>
<td>bootstrap aggregation of C3NET</td>
<td>bc3net</td>
</tr>
</tbody>
</table>

\(^1\) Available from the authors.
Multiple hypothesis testing

There are two types of methods providing either,

- **p-value** for each edge
- **score** for each edge

If p-value, apply multiple testing correction.

Problem:

- high correlation
Part IV

Evaluation measures
Overview of part V

- difficulty of assessing the performance of an inference method
- different types of error measures:
  - general statistical measures
  - oncology-based measures
  - network-based measures
General statistical measures
Ontology-based measures
Network-based measures

- cell type
- biological data
- true network
- inferred network
- evaluation
- reference network
- literature & databases
General statistical measures
Ontology-based measures
Network-based measures

simulated cell type -> simulated data

true network -- evaluation --> inferred network
Different evaluation measures:

- General statistical measures
- Ontology-based measures
- Network-based measures
General statistical measures

sensitivity = \frac{TP}{TP+FN}

specificity = \frac{TN}{TN+FP}

complementary sensitivity = 1 - \text{sensitivity} = \frac{FP}{TN+FP}

precision = P = \frac{TP}{TP+FP}

recall = R = \text{sensitivity} = \frac{TP}{TP+TN}

accuracy = \frac{TP+TN}{TP+TN+FP+FN}

Obtained by comparison of a inferred network with the true network underlying the data.
The first measure is the area under the curve for the receiver operator characteristics (AUC-ROC) [20]. The ROC curve represents the sensitivity as function of the complementary specificity obtained by using various threshold values \( \theta \in \Theta \), instead of one specific, the algorithm depends on. This leads to a \( \theta \)-dependence of all quantities listed above and, hence, allows to obtain a functional behavior among these measures. The second measure is the area under the precision-recall curve (AUC-PR), obtained similarly as AUC-ROC, and the third is the F-score,

\[
F = 2 \frac{PR}{P + R}.
\]

\( F \) assumes values in \([0, 1]\).
- \( F = 0 \) \(\implies\) very bad (worst)
- \( F = 1 \) \(\implies\) very good (best)
Another strategy to assess the performance of an inference method is based on the biological relevance of the inferred network. In general, it is assumed that genes in a gene expression network are preferentially linked to genes involved in similar biological processes [48]. There are several publicly available resources of biological knowledge which can be used to test whether this holds true, for example, the Gene Ontology database (GO) [4] or KEGG [26]. There are also several curated organism-specific knowledge databases, such as RegulonDB (E.coli) [22] and MIPS (yeast) [35]. Functional congruence of clusters of co-expressed genes is a popular validation measure for clustering algorithms [13] and in principle can be used for the validation of inferred networks.
Network-based measures

Problems with general statistical measures:

- global assessment of the whole network
  - Are interactions of TFs better to infer than others?
  - Is a three-gene forward motif more difficult to infer?
- F-score is a random variable
  - mean value of F-score
  - variance of F-score
  - distribution of F-score
Network-based measures

- General statistical measures
- Ontology-based measures
- Network-based measures

Diagram:

- Experiment
- Bootstrap
- Simulation

$G$, $D_1(G)$, $D_2(G)$, $D_b(G)$, $G_1$, $G_2$, $G_p$, $G_b$

- Ensemble data
- Estimated networks

Probabilistic network
General statistical measures
Ontology-based measures
Network-based measures

probabilistic network

weight \( i-j \) = \( \frac{\text{number of times edge } i-j \text{ is present in } \{G_i\}}{b} \)

\( G_1 \)

\( G_2 \)

\( G_b \)

\( G_P \)
General statistical measures
Ontology-based measures
Network-based measures

TPR of \(i-j\) = \(\frac{\text{number of times edge } i-j \text{ is present in } \{G_i\}}{b}\)

TNR of \(i-k\) = \(\frac{\text{number of times no edge } i-k \text{ is present in } \{G_i\}}{b}\)

probabilistic network

knowledge about TP and TN edges

true network

\(G\)
Part V

Data
Different origin of data:

- Biological data
- Simulated data
Biological data
Figure: Central dogma of molecular biology.
Different types of data:

- observational data
- experimental data (intervention, perturbation)

Frequently, only observational data are available because experimental data cannot be collected due to ethic reasons (especially involving humans).
Different types of data:

- observational data
- experimental data (intervention, perturbation)

Frequently, only observational data are available because experimental data cannot be collected due to ethic reasons (especially involving humans).
Simulated data
Simplified kinetic equations (ODE):

\[
\frac{d \text{mRNA}_i}{dt} = S_i(R) - D_i^{\text{mRNA}} \text{mRNA}_i
\]

\[
\frac{d \text{prot}_i}{dt} = T_i^P \text{mRNA}_i - D_i^P \text{prot}_i
\]
Software packages to simulate expression data:

<table>
<thead>
<tr>
<th>name/publication</th>
<th>special features</th>
</tr>
</thead>
<tbody>
<tr>
<td>netsim [16]</td>
<td>R package</td>
</tr>
<tr>
<td>GNW [32]</td>
<td>various dynamic models</td>
</tr>
<tr>
<td>GRENDL [25]</td>
<td>use kinetic parameters from yeast</td>
</tr>
<tr>
<td>SGNSim [39]</td>
<td>stochastic simulation algorithm</td>
</tr>
<tr>
<td>SynTRen [46]</td>
<td>steady-state</td>
</tr>
</tbody>
</table>
Part VI

Practical examples
Overview of part VI

- discussing several practical examples, possible problems and difficulties
Getting started
Suppose: You have 300 chips of expression data with > 10,000 genes.
How to check if the inferred network for a large-scale expression data set makes sense?

Not a good start!
Better: Start with simulations to assess the implemented algorithm.
Getting started

Suppose: You have 300 chips of expression data with > 10,000 genes.

How to check if the inferred network for a large-scale expression data set makes sense?

Not a good start!

Better: Start with simulations to assess the implemented algorithm.
library("igraph"); library("netsim"); library("minet"); library("c3net")

N <- 50; prob <- 0.035
g <- erdos.renyi.game(n=N, p=prob, directed = TRUE, loops = TRUE)
W <- get.adjacency(g, type="both")
ind <- which(W != 0, arr.ind = TRUE)

for(i in 1:dim(ind)[1]){
  if(runif(1) < 0.5) W[ind[i,1], ind[i,2]] <- -1
}

T <- 100; E <- 300
dnew <- matrix(0, nrow = N, ncol = E)
for(i in 1:E){
  aux <- simulateprofiles(weights=W, act.fun="sigmoidal", method="lsoda", times=c(0,1,T))
  dnew[,i] <- aux[,3]
}

net <- c3(build.mim(discretize(t(dnew), disc="equalwidth" ), estimator="mi.mm" ))
res <- validate(net, abs(W))
P <- res[,2]/(res[,2] + res[,3])
R <- res[,2]/(res[,2] + res[,4])
F <- 2*P*R/(P+R)
mF <- max(F, na.rm = TRUE)
library("igraph"); library("netsim"); library("minet"); library("c3net");

N <- 50; prob <- 0.035
g <- erdos.renyi.game(n=N, p=prob, directed = TRUE, loops = TRUE)
W <- get.adjacency(g, type="both")
ind <- which(W != 0, arr.ind = TRUE)

for(i in 1:dim(ind)[1]){
  if(runif(1) < 0.5) W[ind[i,1], ind[i,2]] <- -1
}

T <- 100; E <- 300
dnew <- matrix(0, nrow = N, ncol = E)
for(i in 1:E){
  aux <- simulateprofiles(weights=W, act.fun="sigmoidal", method="lsoda", times =c(0,1,T))
  dnew[i] <- aux[,3]
}

net <- c3(build.mim(discretize(t(dnew), disc="equalwidth" ), estimator="mi.mm" ))
res <- validate( net, abs(W))

P <- res[,2]/(res[,2] + res[,3])
R <- res[,2]/(res[,2] + res[,4])
F <- 2*P*R/(P+R)

mF <- max(F, na.rm = TRUE)
Getting started

1. Algorithm

- simulated cell type
- true network

2. Algorithm

- simulated data

3. Algorithm

- evaluation
- inferred network

Frank Emmert-Streib
Gene Regulatory Networks
Visualization of the network with *igraph*:

tkplot(g, vertex.size=3.0, vertex.color = "blue")
library("igraph"); library("netsim"); library("minet"); library("c3net");

N <- 50; prob <- 0.035
g <- erdos.renyi.game(n=N, p=prob, directed = TRUE, loops = TRUE)
W <- get.adjacency(g, type="both")
ind <- which(W != 0, arr.ind = TRUE)

for(i in 1:dim(ind)[1]){
  if(runif(1) < 0.5) W[ind[i,1], ind[i,2]] <- -1
}

T <- 100; E <- 300
dnew <- matrix(0, nrow = N, ncol = E)
for(i in 1:E){
  aux <- simulateprofiles(weights=W, act.fun="sigmoidal", method="lsoda", times =c(0,1,T))
  dnew[i,] <- aux[,3]
}

net <- c3(build.mim(discretize(t(dnew), disc="equalwidth" ), estimator="mi.mm" ))
res <- validate( net, abs(W))

P <- res[,2]/(res[,2] + res[,3])
R <- res[,2]/(res[,2] + res[,4])
F <- 2*P*R/(P+R)
mF <- max(F, na.rm = TRUE)

Result: maximal F-score: 0.43
Question: Is a F-score of 0.43 a high or low value?
Histogram of 100 F-scores obtained from randomized data.
Question: Is a F-score of 0.43 a high or low value?

Using randomized data as input for the inference algorithm gives in average $F_{\text{rand}} \sim 0.12$
Question: Is the result for $F_{rand}$ independent of the underlying (true) network $G$?
Question: Is the result for $F_{rand}$ independent of the underlying (true) network $G$?

Example:

1. network $G$ with 50 genes constructed by $p = 0.015$
2. generate randomized data
3. MRNET

$F_{rand} \sim 0.07$

$\Rightarrow$ the randomized F-score is a function of the network, $F_{rand}(G)$
Let’s take a closer look:

inferred network (part)

```
> net

V1  0.000000000 0.0244688433 0.027914887 0.013969485 0.000000000 0.000000000
V2  0.024468843 0.0000000000 0.000000000 0.048966411 0.000000000 0.023253497
V3  0.027914887 0.0000000000 0.000000000 0.012782872 0.000000000 0.000000000
V4  0.013969485 0.0489664111 0.012782872 0.000000000 0.000000000 0.013892107
V5  0.000000000 0.0000000000 0.000000000 0.000000000 0.000000000 0.000000000
V6  0.0232534968 0.0000000000 0.000000000 0.012782872 0.000000000 0.000000000
V7  0.000000000 0.0000000000 0.089198730 0.021163258 0.000000000 0.012747563
V8  0.015237811 0.0175343827 0.000000000 0.000000000 0.125830242 0.000000000
V9  0.000000000 0.0000000000 0.000000000 0.000000000 0.000000000 0.000000000
V10 0.044672679 0.9857632600 0.984655214 0.040205932 0.000000000 0.000000000
V11 0.057223037 0.0105832890 0.02417762 0.000000000 0.969407882 0.000000000
V12 0.034959606 0.0636157198 0.000000000 0.000000000 0.089818791 0.020046922
V13 0.000000000 0.0287671936 0.000000000 0.009385961 0.061685278 0.000000000
V14 0.000000000 0.000000000 0.000000000 0.000000000 0.24414013 0.000000000
V15 0.000000000 0.0003002626 0.011006357 0.000000000 0.014411812 0.000000000
V16 0.038310232 0.0089231394 0.029520004 0.015764865 0.000000000 0.000000000
V17 0.063909136 0.000000000 0.000000000 0.003270828 0.022446779 0.040262031
V18 0.01314730 0.0303191134 0.019913482 0.010910713 0.000000000 0.000000000
V19 0.000000000 0.000000000 0.000000000 0.013000233 0.012815517 0.000000000
```
Getting started

inferred network (part)

> net

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>0.000000000</td>
<td>0.024468843</td>
<td>0.027914887</td>
<td>0.013969485</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V2</td>
<td>0.024468843</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.048966411</td>
<td>0.000000000</td>
<td>0.023253498</td>
</tr>
<tr>
<td>V3</td>
<td>0.027914887</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.012782872</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V4</td>
<td>0.013969485</td>
<td>0.048966411</td>
<td>0.012782872</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.013892107</td>
</tr>
<tr>
<td>V5</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.002506736</td>
</tr>
<tr>
<td>V6</td>
<td>0.000000000</td>
<td>0.023253498</td>
<td>0.000000000</td>
<td>0.013892107</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V7</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.089198730</td>
<td>0.021163258</td>
<td>0.000000000</td>
<td>0.12747563</td>
</tr>
<tr>
<td>V8</td>
<td>0.015237811</td>
<td>0.017534383</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.125830242</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V9</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.007453967</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V10</td>
<td>0.044672679</td>
<td>0.985763260</td>
<td>0.984655214</td>
<td>0.040205932</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V11</td>
<td>0.057223037</td>
<td>0.010583289</td>
<td>0.002417762</td>
<td>0.000000000</td>
<td>0.969407882</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V12</td>
<td>0.034959606</td>
<td>0.063615719</td>
<td>0.000000000</td>
<td>0.089818791</td>
<td>0.020046922</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V13</td>
<td>0.000000000</td>
<td>0.028767193</td>
<td>0.000000000</td>
<td>0.061685278</td>
<td>0.061685278</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V14</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.024414013</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V15</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.014411812</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V16</td>
<td>0.038310232</td>
<td>0.008923139</td>
<td>0.029520004</td>
<td>0.015764865</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V17</td>
<td>0.063909136</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.022446779</td>
<td>0.040262031</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V18</td>
<td>0.013114730</td>
<td>0.030319113</td>
<td>0.019913482</td>
<td>0.010910713</td>
<td>0.000000000</td>
<td>0.000000000</td>
</tr>
<tr>
<td>V19</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.000000000</td>
<td>0.013000233</td>
<td>0.012815517</td>
<td>0.000000000</td>
</tr>
</tbody>
</table>

apply threshold

network (unweighted, undirected)
apply **various** thresholds

```r
> res <- validate(net, W)
```

<table>
<thead>
<tr>
<th>thresh</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>67</td>
<td>2433</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.02</td>
<td>26</td>
<td>742</td>
<td>1691</td>
<td>41</td>
</tr>
<tr>
<td>0.04</td>
<td>14</td>
<td>394</td>
<td>2039</td>
<td>53</td>
</tr>
<tr>
<td>0.06</td>
<td>8</td>
<td>216</td>
<td>2217</td>
<td>59</td>
</tr>
<tr>
<td>0.08</td>
<td>4</td>
<td>118</td>
<td>2315</td>
<td>63</td>
</tr>
</tbody>
</table>

**various** networks (unweighted, undirected)

error measures, e.g., F-score

\[
P = \frac{TP}{TP + FP} \\
R = \frac{TP}{TP + FN} \\
F = 2 \frac{PR}{P + R}
\]
apply **various** thresholds

```
> res <- validate( net, W)
```

<table>
<thead>
<tr>
<th>thresh</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>67</td>
<td>2433</td>
<td>0</td>
<td>0.0522</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>26</td>
<td>742</td>
<td>1691</td>
<td>0.0622</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
<td>14</td>
<td>394</td>
<td>2039</td>
<td>0.0589</td>
</tr>
<tr>
<td>4</td>
<td>0.06</td>
<td>8</td>
<td>216</td>
<td>2217</td>
<td>0.0549</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>4</td>
<td>118</td>
<td>2315</td>
<td>0.0423</td>
</tr>
</tbody>
</table>

**various** networks (unweighted, undirected)

error measures, e.g., F-score

\[
P = \frac{TP}{(TP+FP)} \quad F=2 \frac{PR}{(P+R)}\]

\[
R = \frac{TP}{(TP+FN)}
\]
apply *various* thresholds

```r
> res <- validate( net, W)
```

<table>
<thead>
<tr>
<th>thresh</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>2433</td>
<td>0</td>
<td>0</td>
<td>0.0522</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>742</td>
<td>1691</td>
<td>41</td>
<td>0.0622</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>394</td>
<td>2039</td>
<td>53</td>
<td>0.0589</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>216</td>
<td>2217</td>
<td>59</td>
<td>0.0549</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>118</td>
<td>2315</td>
<td>63</td>
<td>0.0423</td>
</tr>
</tbody>
</table>

*various* networks (unweighted, undirected)

maximum F-score

error measures, e.g., F-score

\[
P = \frac{TP}{TP + FP} \quad \quad F = \frac{2 \cdot P \cdot R}{P + R}
\]

\[
R = \frac{TP}{TP + FN}
\]
apply various thresholds

\[
\begin{array}{cccccc}
\text{thresh} & \text{TP} & \text{FP} & \text{TN} & \text{FN} & \text{F-score} \\
1 & 0.00 & 67 & 2433 & 0 & 0 & 0.0522 \\
2 & 0.02 & 26 & 742 & 1691 & 41 & 0.0622 \\
3 & 0.04 & 14 & 394 & 2039 & 53 & 0.0589 \\
4 & 0.06 & 8 & 216 & 2217 & 59 & 0.0549 \\
5 & 0.08 & 4 & 118 & 2315 & 63 & 0.0423 \\
\end{array}
\]

maximum F-score

error measures, e.g., F-score

\[
P = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

\[
R = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

\[
F = \frac{2 \times P \times R}{P + R}
\]

Frank Emmert-Streib

Gene Regulatory Networks
assumption: reference network is available
apply various thresholds

```r
> res <- validate(net, W)
```

<table>
<thead>
<tr>
<th>thrsh</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>67</td>
<td>2433</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>26</td>
<td>742</td>
<td>1691</td>
<td>0.0622</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
<td>14</td>
<td>394</td>
<td>2039</td>
<td>0.0589</td>
</tr>
<tr>
<td>4</td>
<td>0.06</td>
<td>8</td>
<td>216</td>
<td>2217</td>
<td>0.0549</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>4</td>
<td>118</td>
<td>2315</td>
<td>0.0423</td>
</tr>
</tbody>
</table>

various networks (unweighted, undirected)

optimal threshold

maximum F-score

error measures, e.g., F-score

\[
P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F = 2 \frac{PR}{P + R}
\]
Part VII

Applications
Comparison of Methods
Figure: Boxplots for F-scores. A subnetwork of the Yeast GRN with 100 genes is used for the simulations. Ensemble size is $N = 300$ [1].
**Figure**: Boxplots for F-scores. A subnetwork of the TRN of *E. coli* is used for the simulations. Sample size is 1000 and ensemble size is $N = 300$. 
Figure: Subnetwork of yeast consisting of 100 genes, sample size is 200.
### Table: Summary of motif statistics for ARACNE, CLR, MRNET and RN [2].

<table>
<thead>
<tr>
<th>measure/motif type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ARA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#m</td>
<td>40</td>
<td>171</td>
<td>446</td>
<td>10</td>
</tr>
<tr>
<td>(\bar{p})</td>
<td>0.591</td>
<td>0.352</td>
<td>0.530</td>
<td>0.156</td>
</tr>
<tr>
<td>(\sigma(\bar{p}))</td>
<td>0.15</td>
<td>0.04</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>CLR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#m</td>
<td>40</td>
<td>171</td>
<td>446</td>
<td>10</td>
</tr>
<tr>
<td>(\bar{p})</td>
<td>0.506</td>
<td>0.378</td>
<td>0.480</td>
<td>0.171</td>
</tr>
<tr>
<td>(\sigma(\bar{p}))</td>
<td>0.131</td>
<td>0.072</td>
<td>0.137</td>
<td>0.194</td>
</tr>
<tr>
<td><strong>MR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#m</td>
<td>40</td>
<td>171</td>
<td>446</td>
<td>10</td>
</tr>
<tr>
<td>(\bar{p})</td>
<td>0.568</td>
<td>0.326</td>
<td>0.582</td>
<td>0.176</td>
</tr>
<tr>
<td>(\sigma(\bar{p}))</td>
<td>0.1576</td>
<td>0.0083</td>
<td>0.2237</td>
<td>0.1434</td>
</tr>
<tr>
<td><strong>RN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#m</td>
<td>40</td>
<td>171</td>
<td>446</td>
<td>10</td>
</tr>
<tr>
<td>(\bar{p})</td>
<td>0.511</td>
<td>0.321</td>
<td>0.515</td>
<td>0.151</td>
</tr>
<tr>
<td>(\sigma(\bar{p}))</td>
<td>0.121</td>
<td>0.010</td>
<td>0.152</td>
<td>0.112</td>
</tr>
<tr>
<td><strong>CLR (EC)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#m</td>
<td>2105</td>
<td>321896</td>
<td>1315</td>
<td>997</td>
</tr>
<tr>
<td>(\bar{p})</td>
<td>0.3625</td>
<td>0.3353</td>
<td>0.3355</td>
<td>0.0558</td>
</tr>
<tr>
<td>(\sigma(\bar{p}))</td>
<td>0.103</td>
<td>0.026</td>
<td>0.031</td>
<td>0.168</td>
</tr>
</tbody>
</table>

![Gene Regulatory Network Diagrams](image_url)
B cell lymphoma
Case study: B-cell lymphoma

Comparison of Methods
B cell lymphoma

B cell dataset and the analysis of MYC

- ~ 344 samples (GSE2350)
- BCLL, DLBLL, FL, BL, MCL, MM
- Experimental and observational data
- ARACNE network ~ 6000 genes with ~ 129,000 interactions
- 56 genes directly connected to MYC (among 5% largest hubs)
- ChiP analysis for MYC in Ramos Burkitt lymphoma cell line

344 samples Affymetrix hgu95a, hgu95a_v2
RMA and quantile normalization (Irizarry et al. 2003)
median expression for entrez gene identifier with multiple probesets
9,684 genes
C3NET (Altay et al. 2010)
C3NET network components inferred from the B-cell lymphoma dataset

Total 463 components (≥2 genes), 5% largest > 100 genes (25 components)
The giant connected component of the B-cell gene network

884 genes with 883 edges
## Comparison of Methods

### B cell lymphoma

**Cellular component GO enrichment analysis of the giant connected component**

<table>
<thead>
<tr>
<th>GO.ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016021</td>
<td>integral to membrane</td>
<td>2201</td>
<td>282</td>
<td>193.28</td>
<td>2.2e-14</td>
</tr>
<tr>
<td>GO:0031224</td>
<td>intrinsic to membrane</td>
<td>2260</td>
<td>285</td>
<td>198.46</td>
<td>1.2e-13</td>
</tr>
<tr>
<td>GO:0044459</td>
<td>plasma membrane part</td>
<td>1388</td>
<td>195</td>
<td>121.89</td>
<td>4e-13</td>
</tr>
<tr>
<td>GO:0044425</td>
<td>membrane part</td>
<td>2739</td>
<td>327</td>
<td>240.53</td>
<td>1.1e-12</td>
</tr>
<tr>
<td>GO:0005886</td>
<td>plasma membrane</td>
<td>2084</td>
<td>264</td>
<td>183.01</td>
<td>1.3e-12</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>integral to plasma membrane</td>
<td>900</td>
<td>131</td>
<td>79.03</td>
<td>9e-10</td>
</tr>
<tr>
<td>GO:0031226</td>
<td>intrinsic to plasma membrane</td>
<td>915</td>
<td>132</td>
<td>80.35</td>
<td>1.4e-09</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>membrane</td>
<td>3462</td>
<td>372</td>
<td>304.02</td>
<td>4.3e-08</td>
</tr>
<tr>
<td>GO:0034702</td>
<td>ion channel complex</td>
<td>125</td>
<td>31</td>
<td>10.98</td>
<td>6.7e-08</td>
</tr>
<tr>
<td>GO:0005576</td>
<td>extracellular region</td>
<td>1077</td>
<td>136</td>
<td>94.58</td>
<td>3e-06</td>
</tr>
<tr>
<td>GO:0034703</td>
<td>cation channel complex</td>
<td>80</td>
<td>21</td>
<td>7.03</td>
<td>3.4e-06</td>
</tr>
<tr>
<td>GO:0034704</td>
<td>calcium channel complex</td>
<td>22</td>
<td>10</td>
<td>1.93</td>
<td>6.2e-06</td>
</tr>
<tr>
<td>GO:0005903</td>
<td>brush border</td>
<td>31</td>
<td>11</td>
<td>2.72</td>
<td>3.6e-05</td>
</tr>
<tr>
<td>GO:0005891</td>
<td>voltage-gated calcium channel com-</td>
<td>17</td>
<td>8</td>
<td>1.49</td>
<td>4e-05</td>
</tr>
<tr>
<td></td>
<td>plex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0030054</td>
<td>cell junction</td>
<td>327</td>
<td>48</td>
<td>28.72</td>
<td>0.00024</td>
</tr>
</tbody>
</table>

**top 15 Cellular Component terms**
Functional organization of the network components

The largest network components are centered at the cell membrane.
Comparison of Methods
B cell lymphoma

Core structure of the gene network in B-cell lymphoma

B-cell lymphoma C3NET ~ 9,221 edges (total)

Basso (2005), ARACNE ~ 129,000 edges (total)
Summary

- C3NET extracts the core gene network from large-scale gene expression data
- The largest components of the gene network are centered at the periphery of the cell
- **Functional clustering**, genes with similar functional roles are likely to be connected

Part VIII

Summary and Literature
Summary and Literature

This tutorial provided:
- motivation for inferring regulatory networks
- interpretation of gene (regulatory) networks
- overview of network inference methods
- overview of evaluation methods
- discussion of different data types: biological/simulated data
- visualization of gene networks
- some applications to large-scale data sets
Summary and Literature

This tutorial provided:
- motivation for inferring regulatory networks
- interpretation of gene (regulatory) networks
- overview of network inference methods
- overview of evaluation methods
- discussion of different data types: biological/simulated data
- visualization of gene networks
- some applications to large-scale data sets
Review article

- Statistical inference and reverse engineering of gene regulatory networks from observational expression data

Extended tutorial

- Practical introduction to the inference of regulatory networks: toward a systems understanding of cells
  Frank Emmert-Streib
  19th August 2012 at ICSB in Toronto, Canada.
Availability of the Tutorial

- Web: www.bio-complexity.com
- Questions: Email - v@bio-complexity.com
Acknowledgements:

postdoc
Ricardo de Matos Simoes

PhD student
Shailesh Tripathi

former postdoc
Gökmen Altay - now University of Cambridge

Funding:

CCRCB Centre for Cancer Research & Cell Biology

Department for Employment and Learning

EPSRC
They (the students) need to learn how to make the computer do what they want them to do...
If they ever get to the frontier of science there will be no package.

David Botstein, Lewis-Sigler Institute for Integrative Genomics, Princeton University (2010)
http://www.youtube.com/watch?v=qGHvVdBpun0&feature=related
http://www.r-project.org/


Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements.

Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks.

Rank-based edge reconstruction for scale-free genetic regulatory networks.

Learning from Data: Artificial Intelligence and Statistics V, chapter Learning Bayesian networks is NP-complete, pages 121–130.
Springer, 1996.

A map of human cancer signaling.

Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.


Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network.


A protein interaction map of drosophila melanogaster.

Modelling non-stationary gene regulatory processes with a non-homogeneous bayesian network and the allocation sampler.

Benchmarking regulatory network reconstruction with grendel.
KEGG: Kyoto Encyclopedia of Genes and Genomes.

[27] P. Karp, M. Riley, S. Paley, A. Pellegrini-Toole, and M. Krummenacker.
Eco cyc: encyclopedia of escherichia coli genes and metabolism.

Wnt signaling pathway and stem cell signaling network.

[29] T. I. Lee and et al.
Transcriptional regulatory networks in saccharomyces cerevisiae.

Gene regulatory network reconstruction using conditional mutual information.

Human metabolic network reconstruction and its impact on drug discovery and development.

Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods.
Journal of Computational Biology, 16(2), 2009.
ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.  

[34] H. McAdams and L. Shapiro.  
A bacterial cell-cycle regulatory network operating in time and space.  

MIPS: a database for genomes and protein sequences.  

minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information.  

A comprehensive map of the toll-like receptor signaling network.  

[38] P. Oliveri and E. Davidson.  
Gene regulatory network controlling embryonic specification in the sea urchin.  

Sgn sim, a stochastic genetic networks simulator.  
[40] T. Sauka-Spengler and M. Bronner-Fraser. 
A gene regulatory network orchestrates neural crest formation. 

An empirical bayes approach to inferring large-scale gene association networks. 

[42] B. Shipley. 
*Cause and Correlation in Biology.* 

Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. 

A human protein-protein interaction network: a resource for annotating the proteome. 

A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. 


