Functional relationships between genes associated with differentiation potential of aged myogenic progenitors

*Correspondence: Radhakrishnan Nagarajan (radhakrishnan_nagarajan@uams.edu)

1 Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA
2 Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
3 Department of Statistical Sciences, University of Padova, Padova, Italy
4 College of Public Health, University of Arkansas for Medical Sciences, Little Rock, AR, USA
5 Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR, USA
6 College of Health Sciences, University of Kentucky, Lexington, KY, USA

Available from:
Determining Statistically Significant Functional Relationships
The Problem

- Bayesian networks are often used to model the relationships among the components of a biological or natural phenomenon, such as in Holmes [3] and Neapolitan [10].
- In Friedman et al. [1] and Friedman et al. [2] statistically significant functional relationships (FRs) were chosen as those whose confidence was greater than a pre-defined threshold.
- Confidence was defined as the frequency of a given FR across the Bayesian networks learned from nonparametric bootstrap samples.
- The value of the threshold has a dramatic impact on the conclusions, and is especially challenging for small sample sizes – see for example Husmeier [4].
Determining Statistically Significant Functional Relationships

## Estimating the Confidence Threshold

1. Generate a bootstrap sample $X^r_{m \times n}$ from the original data set $X_{m \times n}$ and learn the structure of the Bayesian network from $X^r_{m \times n}$. Determine the corresponding PDAG $\Pi^r$.

2. Generate $X^p_{m \times n}$ by randomly permuting the values in each column of $X_{m \times n}$ and learn the structure of the Bayesian network from $X^p_{m \times n}$. Determine the corresponding PDAG $\Pi^p$.

3. Repeat steps 1 and 2 $g = 1, \ldots, n_s$ times to get the PDAGs $\Pi^r_g$ and $\Pi^p_g$.

4. Determine the confidence of the arcs $X_i \rightarrow X_j$, $i \neq j$ in the resampled networks $\Pi^r_g$, $\{ f^r_{ij} \}$, and in the permuted networks $\Pi^p_g$, $\{ f^p_{ij} \}$.

5. An arc $X_i \rightarrow X_j$ is deemed significant if $f^r_{ij} > f^p_{gh}$, $g, h = 1, \ldots, n$, $g \neq h$. 

Marco Scutari
University of Padova
Determining Statistically Significant Functional Relationships

Estimating the Confidence Threshold

Marco Scutari  
University of Padova
Estimating the Confidence Threshold

\[ \{ f_{ij}^p \} \]
Determining Statistically Significant Functional Relationships

Estimating the Confidence Threshold

noise-floor from the permutations

significant arcs

\{ f_{gh}^p \}

\{ f_{ij}^r \}
Properties of the Estimated Confidence Thresholds

The proposed algorithm is essentially a non-parametric bootstrap that estimates the joint empirical distribution of the arc frequencies from the data and compares it to the null distribution of arc frequencies obtained from the randomly permuted counterpart. Note that:

- the correlation structure of the data is destroyed by the permutation, so the edge frequencies $f_{gh}^p$ essentially represent the noise-floor.

- the use of random permutations does not require additional assumptions on the data since the gene expression measurement across the replicate clones is generated independently.

- inference is exact conditionally on the observed sample – i.e. the tests are invariant to the underlying statistical distribution of the data, which may be partially or completely unknown.
Tests on the ASIA Data Set

The proposed algorithm was first tested on data sampled from the ASIA network using three different structure learning algorithms: PC as implemented by Kalisch and Maechler [5], and GS and IAMB as implemented by Scutari [11, 12].

1. generate the true PDAG of the network, $\Sigma^0$.
2. identify significant arcs $\Sigma^1$ from the given empirical sample using one of the proposed algorithms.
3. identify significant arcs $\Sigma^2$ from the given empirical sample using a pre-defined threshold $\theta = (0.05, 0.25, 0.50, 0.75, 0.95)$.
4. compute true and false positive rates from $(\Sigma^0, \Sigma^1)$ and $(\Sigma^0, \Sigma^2)$. 
The ASIA Data Set

The ASIA network from S. L. Lauritzen and D. J. Spiegelhalter [6].
1. the algorithm indeed has low FPR and high TPR.

2. the algorithm performs considerably better than $\theta = (0.50, 0.75, 0.95)$ for samples of size 5000 and 34 (the sample size of the myogenic data set).

3. performance is comparable in the other cases for sample size 5000, but is still better for sample size 34.

So:

1. it is possible to choose a good value for $\theta$, but it depends on the data and the sample size.

2. it is difficult to pick a good, statistically motivated value of $\theta$ in $[0, 1]$; the proposed algorithm does it automatically in a data-driven way.
Analysis of Osteoprogenitor Differentiation
The probabilistic mechanism underlying osteoprogenitor differentiation was established in Madras et al. [7] using 8 genes (COLL1, OCN, ALP, BSP, FGFR1, PTH1R, PTHrP and PDGFRα) and was also studied using Bayesian networks and a pre-defined threshold in Nagarajan et al. [8].

There are two reasons why we chose to re-investigate this data:

- the experimental design of the osteoprogenitor differentiation is similar to that of myogenic progenitor differentiation.
- using the proposed algorithm over real data shows that it may really identify biologically relevant and novel FRs.
Statistically Significant FRs

Analysis of Osteoprogenitor Differentiation

$n = 34$

Marco Scutari
University of Padova
Analysis of Myogenic Progenitors
transcriptions of regulatory (gene) networks controlling both myogenic and adipogenic differentiation are still under active investigation.

myogenic and adipogenic differentiation pathways are typically considered non-overlapping, but Taylor-Jones et al. [13] has shown that myogenic progenitors from aged mice co-express some aspects of both myogenic and adipogenic gene programs.

their balance is apparently regulated by Wnt signaling according to Vertino et al. [14], but there have been few efforts to understand the interactions between these two networks.
The Experimental Setting

The clonal gene expression data was generated from RNA isolated from 34 clones of myogenic progenitors obtained from 24-months old mice, cultured to confluence and allowed to differentiate for 24 hours. RT–PCR was used to quantify the expression of 12 genes:

- myogenic regulatory factors: Myo-D1, Myogenin and Myf-5.
- adipogenesis-related genes: FoxC2, DDIT3, C/EPB and PPARγ.
- Wnt-related genes: Wnt5a and Lrp5.
- control genes: GAPDH, 18S and B2M.
Statistically Significant FRs

control genes:
GAPDH, 18S, B2M

Marco Scutari
University of Padova
Conclusions and Future Research

• While the FRs identified in the present study may not necessarily represent direct relationships, they clearly establish the orchestration of differentiation pathways in aged myogenic progenitor differentiation and their interaction.

• The proposed resampling approach obviates the need for a pre-defined threshold, and has been shown to work well even at small sample sizes.

• Still missing: multiple testing corrections in the structure learning algorithm to control family-wise error rate and/or false-discovery rate and comparing the network structure obtained on the aged myoblasts to those obtained on adult myoblasts.
Thank you for attending.
<table>
<thead>
<tr>
<th>References</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N. Friedman, M. Goldszmidt, and A. Wyner.</td>
<td>Data Analysis with Bayesian Networks: A Bootstrap Approach.</td>
</tr>
<tr>
<td></td>
<td>N. Friedman, M. Linial, and I. Nachman.</td>
<td>Using Bayesian Networks to Analyze Expression Data.</td>
</tr>
<tr>
<td></td>
<td>Springer-Verlag, 2008.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R package version 0.1-8.</td>
<td></td>
</tr>
</tbody>
</table>


