Co-occupancy networks for histone modifications and chromatin associated proteins

Martin Vingron
MPI for Molecular Genetics

Acknowledgements: Ho-Ryun Chung, Rosa Karlic, Julia Lasserre, Juliane Perner
Co-occupancy Networks

• Can we predict gene expression? Predict from what? Histone modifications?
• Do „things“ occupy DNA together?
• Histone modification networks
• Partial correlation, Gaussian Graphical Models
• Histone modifications plus chromatin modifiers
• Compare to: Gene expression networks, e.g. BNs, $p>>n$ problem
Histone modifications

The figure illustrates nucleosome models and major posttranslational modifications which play essential roles in gene expression regulation and disease processes.
18 Histone acetylations

Combinatorial patterns of histone acetylations and methylations in the human genome

Zhibin Wang¹,⁵, Chongzhi Zang²,⁵, Jeffrey A Rosenfeld³,⁵, Dustin E Schones¹, Artem Barski¹, Suresh Cuddapah¹, Kairong Cui¹, Tae-Young Roh¹, Weiqun Peng², Michael Q Zhang³ & Keji Zhao¹

Gene expression data

Dynamic Regulation of Nucleosome Positioning in the Human Genome

Dustin E. Schones¹,² Kairong Cui¹,² Suresh Cuddapah¹, Tae-Young Roh¹, Artem Barski¹, Zhibin Wang¹, Gang Wei¹ and Keji Zhao¹

All data from a single cell type: human CD4+ T-cells

Control: CD4+ goat IgG and CD4+ rabbit IgG (Wang et al., 2009)
Histone modifications and transcription level

Focus on
- Sum of tags in promoter
- For many histone modifications
- in one and the same cell line.

Data from Keiji Zhao lab.

Results in a data matrix size #promoters x #histone modifications.
Determine the number $N_i$ of tags mapping to $\pm 2000$ base pairs from the TSS

Transform $N_i$ by

$$X_i = \log(N_i + \alpha_i)$$

$\alpha_i$ (pseudocount) is chosen such that the correlation with the expression value is maximal

Standard linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i,$$

$i = 1, \ldots, n$
Histone modifications and transcription level

\[ r = 0.77 \]
Feature Selection

- Identify best three-modification linear models
- Find overrepresented modifications
Human Promoter Classes

N = 4,183
N = 10,619

Mikkelsen et al. (2007) Nature 448, 548
Informative modifications stratified by CpG contents
Correlations among modifications

Correlation between different histone modifications in the promoter regions of 14,802 human genes
Look at co-occurrence of TFBSs in a window on the genome

Thomas Manke (J. Mol. Biol., 2003)
Detection of co-regulating TFs

- non-random binding of a TF pair on hypersensitive sites
- cell-type/tissue specific
- TFs: represented as ranked list of hypersensitive sites ordered by binding affinity
- idea: TF pairs with similar top-ranked lists might co-regulate the transcription of their target genes
Predicted interactions in hematopoietic stem cells

Hematopoiesis
Correlations among modifications

Correlation between different histone modifications in the promoter regions of 14,802 human genes
From correlations to partial correlations (Gaussian Graphical Models)

- Data matrix of dimension $N \times H$
  - $N$: number of genes
  - $H$: number of variables
- For each pair $(h_i, h_j)$ of variables compute the correlation $c_{ij}$ between $r_i$ and $r_j$

This and following slides: Julia Lasserre et al
Partial correlation coefficient for variables i and j:
Regress i and j on the remaining variables. Determine the two sets of residuals after explaining i and j.

Partial correlation between i and j is defined as the correlation between the two vectors of residuals.
Theorem on how to compute partial correlations:

Partial correlation coefficients can be computed from the entries of the inverse of the variance-covariance matrix.

Theorem on the meaning:

Under the assumption that the variables are multivariate Gaussian, the partial correlation $\rho_{X Y \cdot Z}$ is zero if and only if $X$ is conditionally independent from $Y$ given $Z$. 
Example

\[ \mathbf{X} = \mathbf{\Sigma} \]

\[
\begin{array}{cccc}
X_1 & X_2 & X_3 & X_4 \\
\hline
X_1 & 1.0 & 2.0 & 6.0 & 3.0 \\
X_2 & 2.0 & 5.0 & 15.1 & 7.6 \\
X_3 & 6.0 & 15.1 & 46.3 & 23.1 \\
X_4 & 3.0 & 7.6 & 23.1 & 12.6 \\
\end{array}
\]

\[
\mathbf{\Sigma}_{ij} = 0 \quad \iff \quad \text{Independence}
\]

\[
\begin{array}{cccc}
X_1 & X_2 & X_3 & X_4 \\
\hline
X_1 & 5.0 & -2.0 & 0.0 & 0.0 \\
X_2 & -2.0 & 10 & -3.0 & 0.0 \\
X_3 & 0.0 & -3.0 & 1.2 & -0.5 \\
X_4 & 0.0 & 0.0 & -0.5 & 1.0 \\
\end{array}
\]

\[
\Sigma_{ij}^{-1} = 0 \quad \iff \quad \text{Conditional independence}
\]

\[X_1 \sim N(0, 1)\]
\[X_2 \sim N(2X_1 + 1, 1)\]
\[X_3 \sim N(3X_2 - 0.5, 1)\]
\[X_4 \sim N(0.5X_3, 1)\]

Slide by Juliane Perner
CD4+ network
Odds and Ends

- We use rank-sorted data (corresponding to a rank correlation coefficient)
- There is huge number of entries in the vectors for which we compute partial correlation coefficients. Therefore we get tiny p-values.
- Remedy: Choose a p-value cutoff and resample from the promoters.
- Accept edges with more than, say, 70% bootstrap support.
Enter chromatin modifiers ....

Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells.
Oren Ram, Alon Goren, Ido Amit, Noam Shoresh, Nir Yosef, Jason Ernst, Manolis Kellis, Melissa Gymrek, Robbyn Issner, Michael Coyne, Timothy Durham, Xiaolan Zhang, Julie Donaghey, Charles B. Epstein, Aviv Regev, Bradley E. Bernstein
Cell, Vol. 147, No. 7. (23 December 2011), pp. 1628-1639

This and following slides: Juliane Perner et al
Correlations among Histone Modifications plus Chromatin Modifiers
Partial Correlations among Histone Modifications plus Chromatin Modifiers
More network construction …

- We modeled expression from HMs, with subsequent feature selection:

```
HM1
HM2
HM3
HM4
HM5
HM6
```

expression
More network construction: HM-> CM

- Why not model a HMs from CMs?
More network construction: HM-> CM

• Why not model all HMs from CMs?
More network construction: sparse linear regression, elastic net

- Sparse regression replaces feature selection
Sparse linear model (elastic net) explaining Histone Modifications from Chromatin Modifiers
Chromatin-signalling network.

aka LSD1=lysine specific histone demethylase, demethylates mono- and di-methylated H3K4

Verification of two predicted interactions links H4K20me1 to Polycomb-mediated repression.
#nodes >> #conditions?
Not with histone modifications!
Acknowledgements

Ho-Ryun Chung, Rosa Karlic … Linear models, HMs
Julia Lasserre .... Gaussian Graphicial Models
Juliane Perner … HM-CM networks
Sarah Kinkley … validation experiments