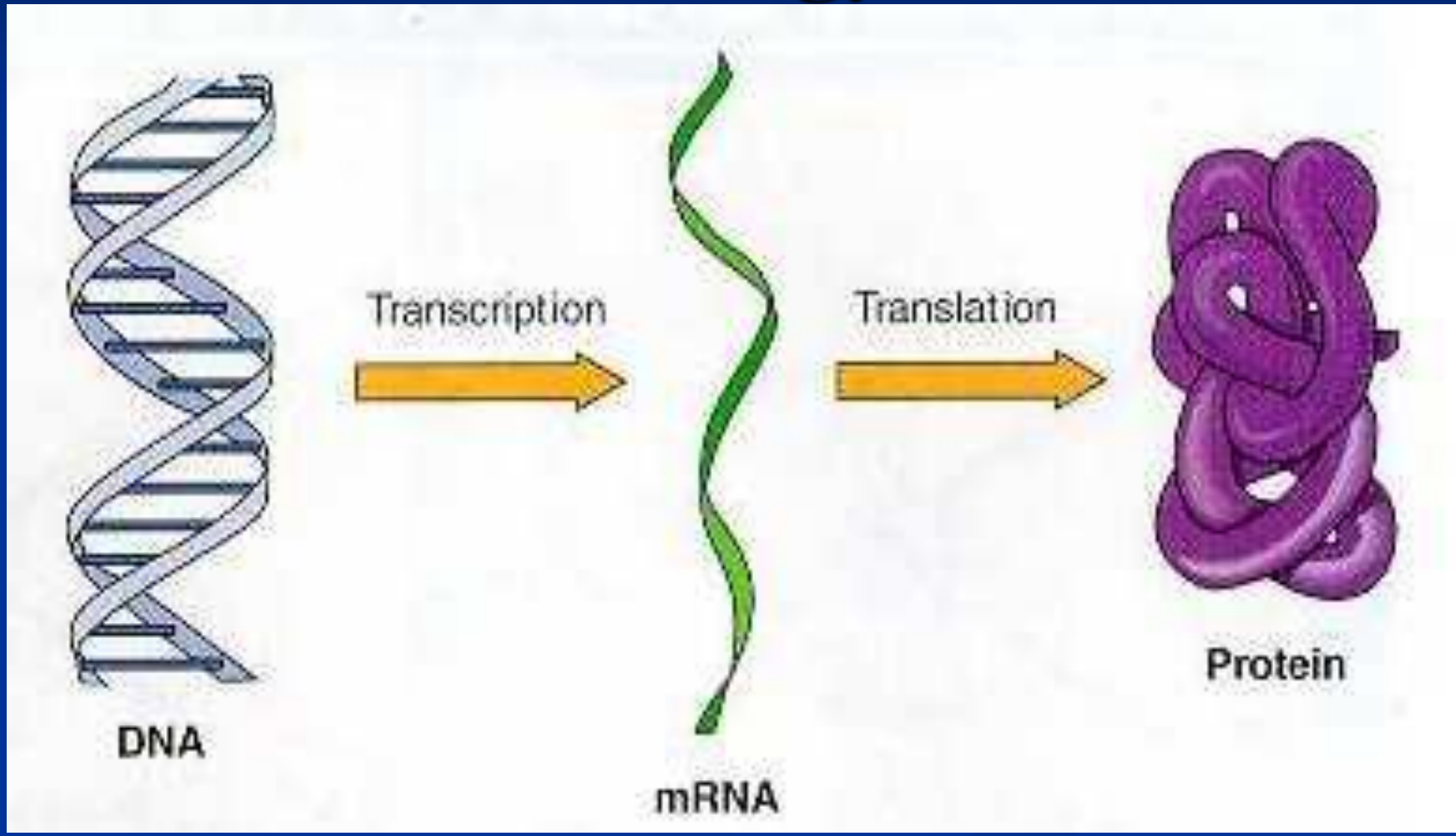
Nonparametric Modeling of Regulatory Network

Ping Ma
Department of Statistics
& Institute for Genomic Biology
University of Illinois Urbana-Champaign

Central Dogma of Molecular Biology



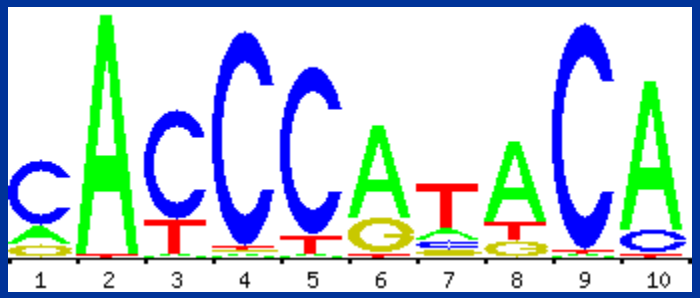
Transcription Regulation

Transcription factors (regulatory proteins) bind to genes, turning on or shutting off their expressions.

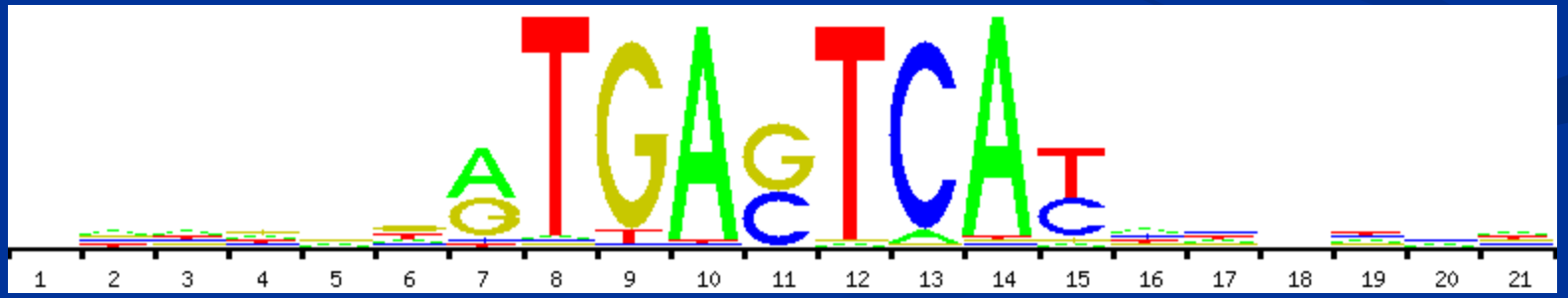
Transcription Factor Binding

- Transcription Factor Binding Motif (TFBM): Common patterns in DNA sequences at transcription factor binding sites.

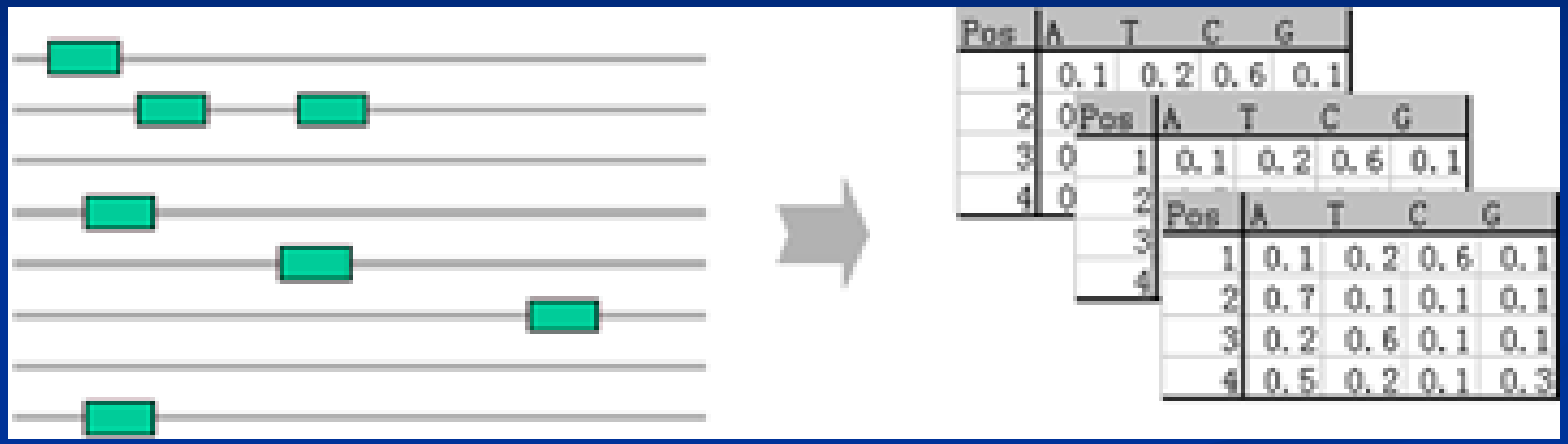
RAP1



GCN4

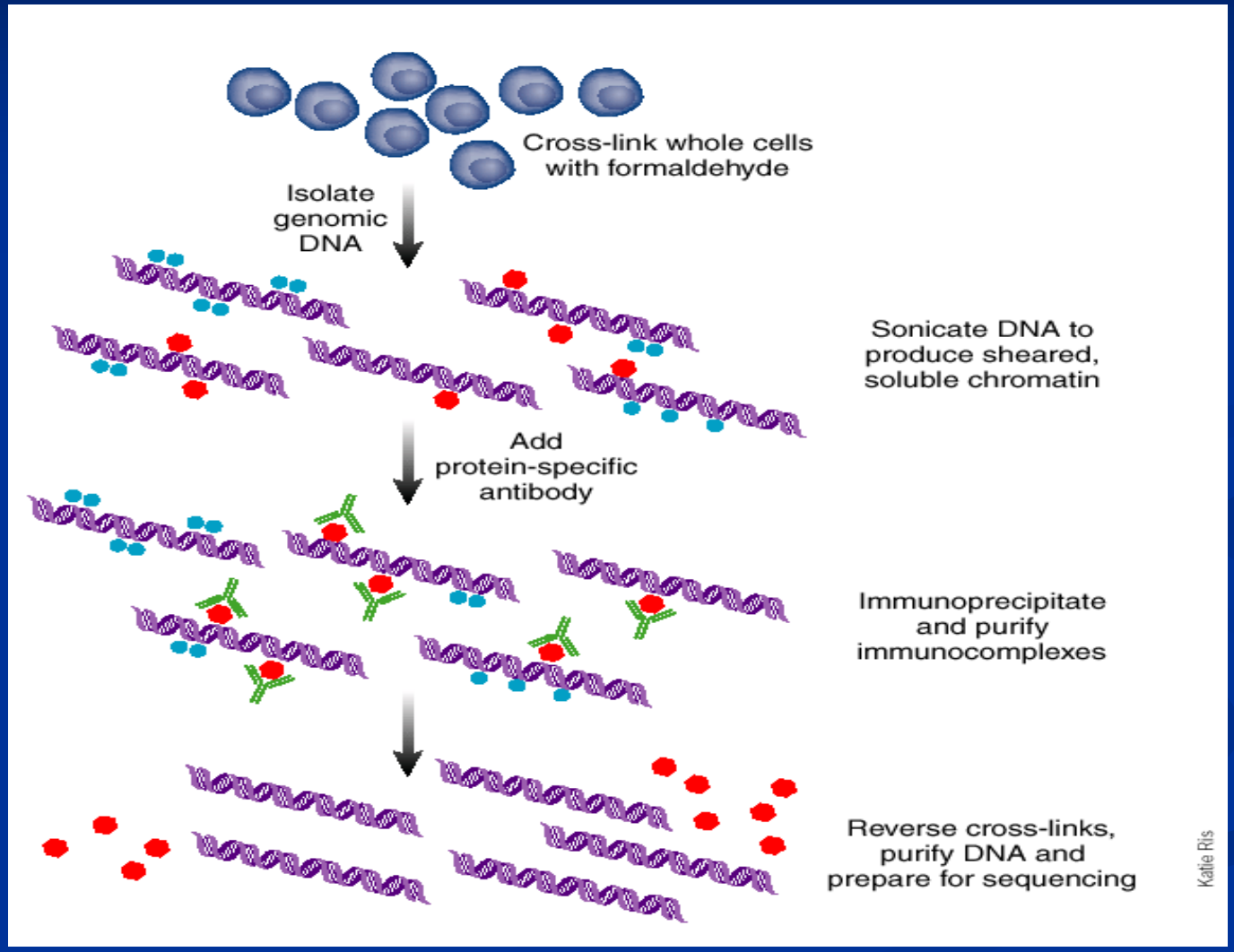


Transcription Factor Binding



Gene	Mf1	Mf2	Mf3	…
gene1	X	X	X	
gene2	X	X	X	
gene3	X	X	X	
…				

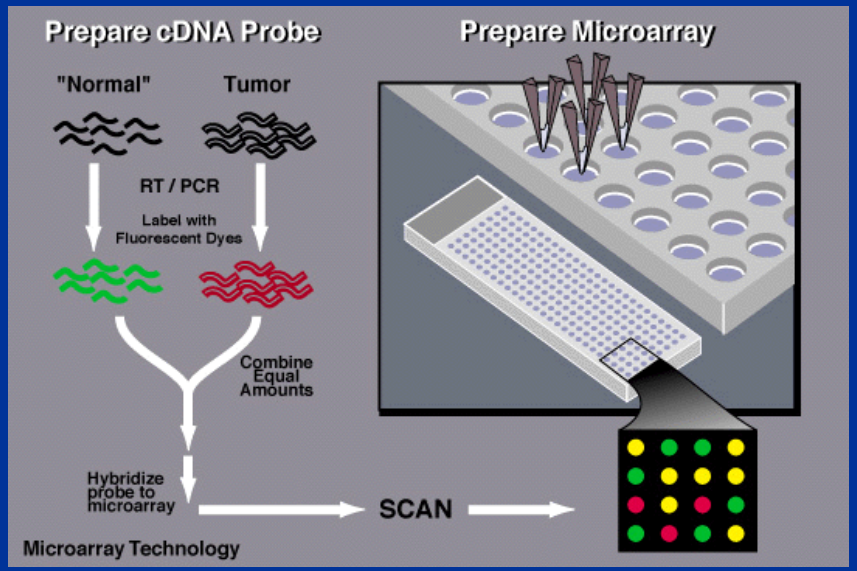
Transcription Factor Binding



Mardis Nat Meth. 2007

Gene Expression

- To quantify the abundance of each transcript
- Two approaches:



Hybridization (Microarray)

Sequence (RNA-Seq)

Linking gene expression with TF binding

- Linear Regression

  Motif Regressor (Conlon et al 2003 PNAS)

  Motif Express (Zamdborg and Ma 2009 NAR)

- Nonlinear Regression

  RSIR (Zhong et al 2005, Bioinformatics)

  Correlation Pursuit (Zhong et al 2012, JRSSB)

Converting Gene Expression to Clusters

- Gene expression is noisy
- Clustering gene expression to get robust clusters
- Linking gene clusters with TF binding data.

  Bayesian Network (Beer and Tavazoie 2004 Cell)

  Proportional Odds Model (Yuan et al 2007 PLoS Comput. Biol.)

Desirable Features

- Flexible function form to link gene expression (clusters) with TF binding
- Integration of new expression data

Our Method

- Gene expression clusters and TF binding

Notation: $y = (y_{\langle 1 \rangle}, \ldots, y_{\langle \Gamma \rangle}) \in \mathcal{Y} = \prod_{\gamma=1}^{\Gamma} \mathcal{Y}_\gamma$, $y_{\langle \gamma \rangle} \in \mathcal{Y}_\gamma$, $x \in \mathcal{X}$.

Task: Estimate $p(y|x)$; it is a special case of conditional density estimation on generic $\mathcal{X} \times \mathcal{Y}$ (Gu 1995, SS), for $\mathcal{Y}$ discrete.

Penalized Likelihood

To estimate $p(y|x) = e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$, for $\eta = \eta_y + \eta_{xy}$, minimize

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(x_i, y_i) - \log \int_{\mathcal{Y}} e^{\eta(x_i,y)}\right\} + \lambda J(\eta),$$

where $J(\eta)$ is roughness functional and λ is smoothing parameter.

Functional ANOVA

- On $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$, for averaging operators $A_x 1 = 1$, $A_y 1 = 1$,

$$\begin{aligned}\eta(z) &= (I - A_x + A_x)(I - A_y + A_y)\eta \\ &= A_x A_y \eta + (I - A_x)A_y\eta + A_x(I - A_y)\eta + (I - A_x)(I - A_y)\eta \\ &= \eta_\emptyset + \eta_x(x) + \eta_y(y) + \eta_{xy}(x, y),\end{aligned}$$

with side conditions $A_x\eta_x = A_x\eta_{xy} = A_y\eta_y = A_y\eta_{xy} = 0$.

- On $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$, with $\eta = \eta_\emptyset + \eta_u = \eta_\emptyset + \eta_x + \eta_y + \eta_{xy}$,

$$p(x, y) = \frac{e^{\eta_u}}{\int_{\mathcal{U}} e^{\eta_u}} = \frac{e^{\eta_x + \eta_y + \eta_{xy}}}{\int_{\mathcal{X}\times\mathcal{Y}} e^{\eta_x + \eta_y + \eta_{xy}}}, \quad p(y|x) = \frac{e^{\eta_y + \eta_{xy}}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{xy}}}.$$

Penalized Likelihood

To estimate $p(y|x) = e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$, for $\eta = \eta_y + \eta_{xy}$, minimize

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(x_i, y_i) - \log \int_{\mathcal{Y}} e^{\eta(x_i,y)}\right\} + \lambda J(\eta),$$

where $J(\eta)$ is roughness functional and λ is smoothing parameter.

- ANOVA structures $\eta = \sum_\beta \eta_\beta$ are built in via tensor product splines, with $J(\eta) = \sum_\beta {\theta_\beta}^{-1} J_\beta(\eta_\beta)$ involving extra smoothing parameters θ_β.
- Smoothing parameters are selected by cross-validation.

Cross-Validation

- $\mathrm{KL}(\eta, \eta_\lambda) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} (\eta - \eta_\lambda) p(y|x) - \log \int_{\mathcal{Y}} e^{\eta} + \log \int_{\mathcal{Y}} e^{\eta_\lambda} \right\} p(x)$.
- Estimate $\int_{\mathcal{X}\times\mathcal{Y}} \eta_\lambda(x,y) p(x,y)$ by $\frac{1}{n}\sum_i \eta_\lambda^{[i]}(x_i, y_i)$.
- $V(\lambda) = -\frac{1}{n}\sum_i \left\{\eta_\lambda(x_i, y_i) - \log \int_{\mathcal{Y}} e^{\eta_\lambda(x_i,y)}\right\} + \frac{1}{n}\sum_i \left(\eta_\lambda - \eta_\lambda^{[i]}\right)(x_i, y_i)$

Inference

- Task: Test $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$.
- KL Projection: Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, minimize $\mathrm{KL}(\hat{\eta}, \eta)$ over $\eta \in \mathcal{H}_0$ to obtain $\tilde{\eta}$, then inspect the "entropy" decomposition,

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c),$$

  where $\eta_c \in \mathcal{H}_0$ is a "baseline" fit; Gu (2004, CJS).
- If $\rho = \mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ is small, one loses little by cutting out $\mathcal{H}_1$.
- One may take $\eta_c = \eta_1 + \cdots + \eta_\Gamma$.

Bayesian Confidence Interval

- A quadratic $J(\eta)$ acts like the (minus) log likelihood of a Gaussian process prior for η.
- Substituting $L(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(x_i, y_i) - \log\int_{\mathcal{Y}} e^{\eta(x_i, y)}\right\}$ by its quadratic approximation $Q_{\hat{\eta}}(\eta)$ at $\hat{\eta}$, $Q_{\hat{\eta}}(\eta) + \lambda J(\eta)$ appears as a Gaussian posterior log likelihood with $E\left[\eta(x, y)\right] = \hat{\eta}(x, y)$.
- CIs for $\eta(x, y)$ have little meaning, as $p(y|x) = e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$.
- y-Contrasts: $\theta(x) = \sum_y c_y \log p(y|x) = \sum_y c_y \eta(x, y)$, for $\sum_y c_y = 0$.
  - The normalizing constant $\int_{\mathcal{Y}} e^{\eta(x,y)}$ cancels out in y-contrasts.
- Based on $Q_{\hat{\eta}}(\eta) + \lambda J(\eta)$, one may calculate $E\left[\theta(x)\right] = \sum_y c_y \hat{\eta}(x, y)$ and $\mathrm{Var}\left[\theta(x)\right]$ to construct Bayesian confidence intervals for $\theta(x)$.

Mixed Effect Models

- For univariate responses, one may use $\zeta_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b}$, $\mathbf{b} \sim N(0, B)$, where $\eta(x)$ is fixed-effect and $\mathbf{z}^T \mathbf{b}$ comprises random effects.
- Given $\int_{\mathcal{Y}} \eta = 0$, specify $\mathbf{b}_y$ satisfying $\int_{\mathcal{Y}} \mathbf{b}_y = 0$, then minimize $-\frac{1}{n} \sum_{i=1}^{n} \left\{ \eta(x_i, y_i) + \mathbf{z}_i^T \mathbf{b}_{y_i} - \log \int_{\mathcal{Y}} e^{\eta(x_i, y) + \mathbf{z}_i^T \mathbf{b}_y} \right\} + \tilde{\mathbf{b}}^T \Sigma \tilde{\mathbf{b}} + \lambda J(\eta)$, where $\tilde{\mathbf{b}}$ comprises components of $\mathbf{b}_y$, with $\tilde{\mathbf{b}} \sim N(0, c\Sigma^+)$.

The Yeast Data

- Yeast Stress Experiment: Gasch et al (2000, Mol. Biol. Cell).
  - Yeast samples were put under environmental stresses such as heat shock, hydrogen peroxide, and amino acid starvation.
  - Gene expressions were measured before and after the application of the stress, and genes are classified as responsive or non-responsive to the stress.
- TFBM Matching Scores: Beer and Tavazoie (2004, Cell).
  - A motif was compared against the upstream 800 base pairs of a gene, and a matching score was calculated.
  - A higher score results from more frequent or better quality matches, indicating more likely binding.
- The data (x_i, y_i) are from $n = 2587$ genes.
  - $y \in \{0,1\}^3$: Responsiveness to 3 environmental stresses.
  - $x \in R^{51}$: Matching scores to 51 TFBMs.

Data Screening: Yeast Data

- Variable Screening: x's are skewed with many 0's, so $\sqrt{x}$ is used.
- For each x-variable, fit $\hat{\eta} = \eta_y + \eta_{xy}$, and project $\hat{\eta}$ to $\tilde{\eta} = \eta_y$, obtaining $\rho = \mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$; η_y, η_{xy} both have 7 terms.

TF	PAC	RRPE	GCN4	RAP1	CAD1	YAP1	···
ρ	0.600	0.446	0.293	0.284	0.255	0.221	···

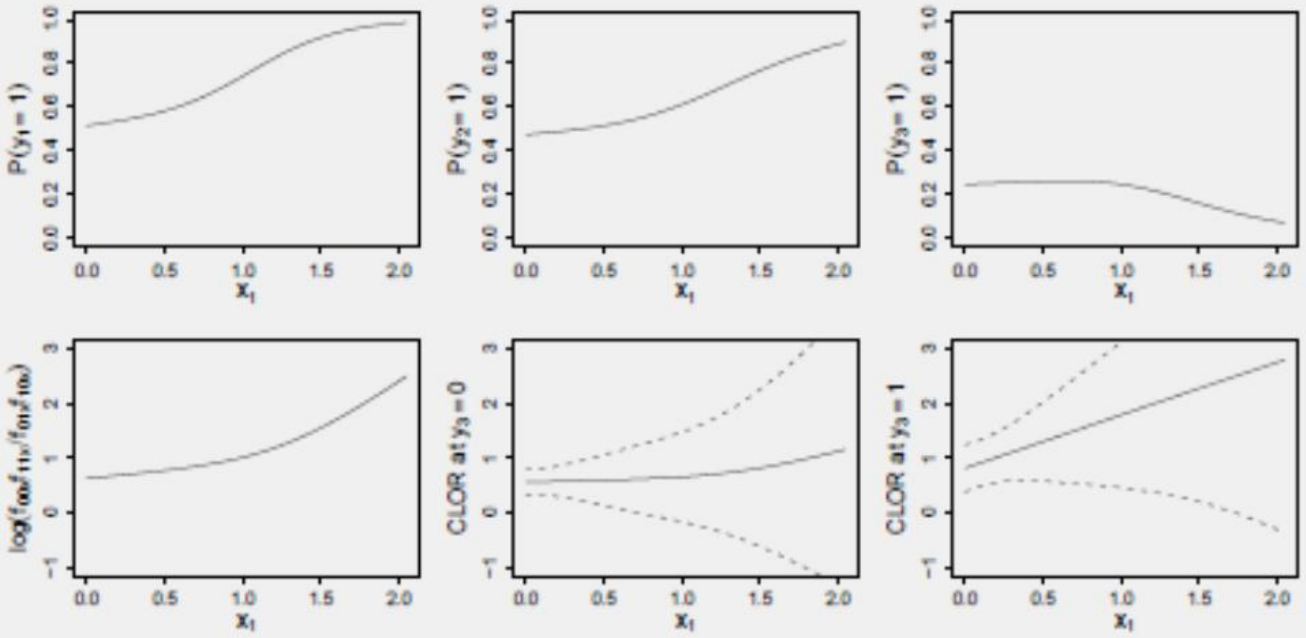
- The top 5 on the list were used in further analysis.
- "Half" of data are at origin (1203/2587), and the "rest" on axes.
- Counts of non-zero $x_{(i)}x_{(j)}$ are listed below.

	PAC	RRPE	GCN4	RAP1	CAD1
PAC	416	211	38	53	68
RRPE		534	59	70	98
GCN4			273	42	49
RAP1				323	53
CAD1					447

Data Analysis: Yeast Data

- Initial Model: 7 terms in η_y, $7 \times 7 = 49$ terms in η_{xy}, where on the x-axis one has $\eta_x = \eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_{12} + \eta_{25}$.
- Final Model: 7 terms in η_y, $7 \times 6 = 42$ terms in η_{xy}, with η_{25} dropped out on the x-axis.

- The effects of PAC are shown below, with the other 4 TF fixed at 0.



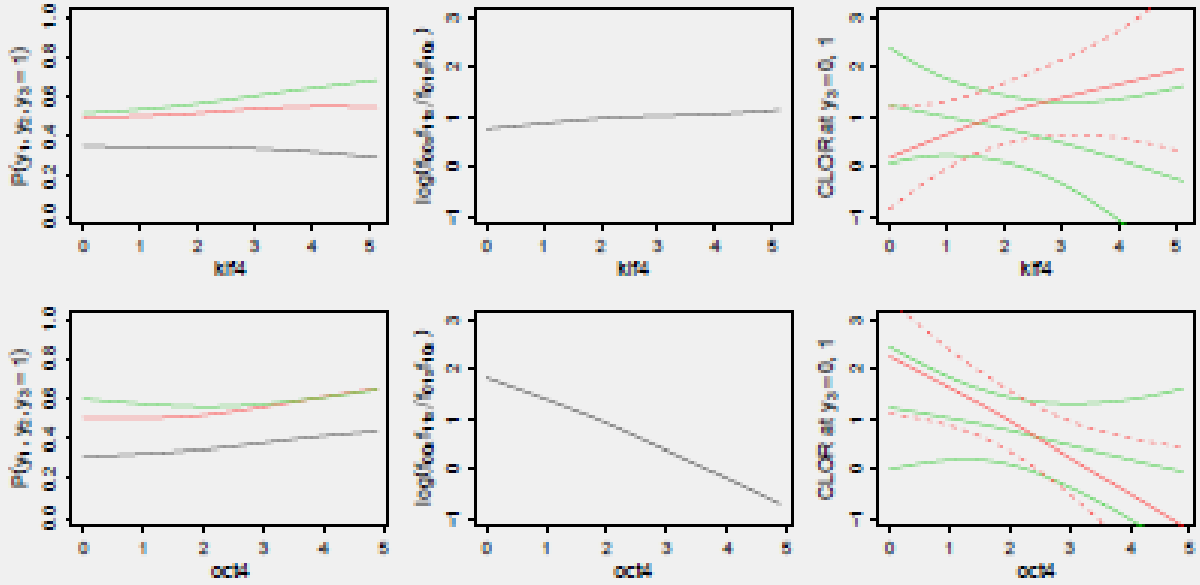
The Stem Cell Data

- Gene expression of mouse embryonic stem cells: Cai et al (2010).
  - Expression levels measured on day 0, 4, 8, 14; $\{4, 8, 14\}$ compared against 0.
  - Genes are tagged as differentially expressed or not, at 3 time points.
- Transcription Factor Association Strength: Ouyang et al (2009).
  - Association strengths between TFs and genes, based on ChIP-seq profiles.
- The data (x_i, y_i) are from $n = 1027$ genes.
  - $y \in \{0, 1\}^3$: Expression at 3 time points.
  - $x \in R^4$: TFAS scores of 4 TFs, NANOG, SOX2, OCT4, and KLF4.
  - NANOG, SOX2, OCT4 regulate pluripotency; KLF4 regulates differentiation.
  - Genes are also clustered as up-regulation and down-regulation.

Data Analysis: Stem Cell Data

- Initial Model: 7 terms in η_y, $7 \times 15 = 105$ terms in $\eta_{x,y}$, where on the x-axis all interactions were included.
- Final Model: 7 terms in η_y, $7 \times 13 = 91$ terms in $\eta_{x,y}$, with terms involving NANOG–SOX2–OCT4 removed.

- "Slices" of the fit are shown below, with the other TFASs fixed at medians.



Left: $P(y_{(1)}, y_{(2)}, y_{(2)} = 1)$. Right: Log odds ratio of $y_{(1)}$, $y_{(2)}$ at $y_{(3)} = 0, 1$.

Software

- R package gss

http://cran.r-project.org/web/packages/gss/

Joint work with Chong Gu