PUMA: Propagation of Uncertainty in Microarray Analysis

Low Level and High Level Processing of Microarrays with Probabilistic Models

Neil Lawrence
Department of Computer Science
University of Sheffield
Magnus Rattray
School of Computer Science
University of Manchester

August 2, 2006

Outline

1. Microarray Processing
   - Affymetrix GeneChip Arrays
   - Detecting Differential Gene Expression with PPLR
   - Tidying up Profiles with Probabilistic PCA
2. Transcription Factors
   - ChIP-microarray and Transcription Factor Activities
   - Transcription Factor Concentrations
   - From Simple to Complex Models
3. Conclusions

Online Resources

All source code and slides are available online

- This talk available from my home page (see talks link on side).
- Project main page (with links to software)
  - http://bioinf.man.ac.uk/resources/puma/.
- Additional project homepage
  - http://www.dcs.shef.ac.uk/~neil/projects/pipeline/.

PUMA Project Outline

Noise Problems in Microarrays

- Project was motivated by the fact that microarray data is very noisy.
- The aim of the project is to:
  - Assess the level of noise in the estimated gene expression.
  - Propagate the noise through downstream analysis.
- Personnel:
  - **Investigators**: Neil Lawrence (Sheffield), Magnus Rattray (Manchester)
  - **Fellows/Post-docs**: Marta Milo (Sheffield), Guido Sanguinetti (Sheffield)
  - **PhD Students**: Xuejun Liu (Manchester), Richard Pearson (Manchester)

Central Dogma

DNA →mRNA →Protein

- Every cell has the same DNA.
- Cells produce different proteins (building blocks of life).
- Level of mRNA produced is known as *gene expression*.
- Has a downstream effect on level of Protein produced.
- Gene expression is controlled by *Transcription factors*.
- Transcription factors themselves are proteins.
    - Feedbacks in these systems lead to gene networks.

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

- Affymetrix arrays are a technology for measuring level of mRNA.
- PM (perfect match) probes match the gene sequence.
- MM (mismatch) probes have wrong middle base.
- MM designed to measure non-specific binding.
- Approx 10,000 probe-sets per chip.

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

Figure: Affymetrix arrays for human and mouse (image from Wikimedia Commons under GFDL).

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

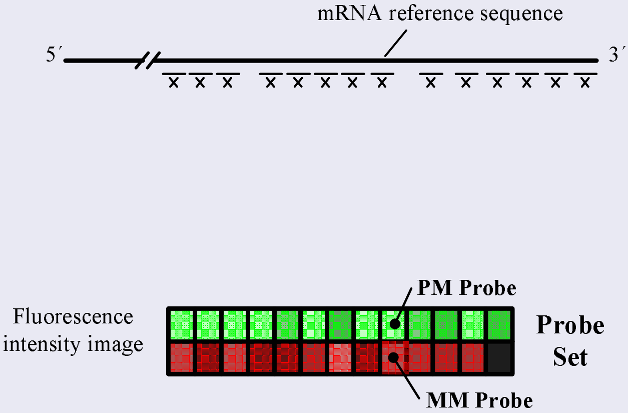


Figure: Affymetrix array schematic

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

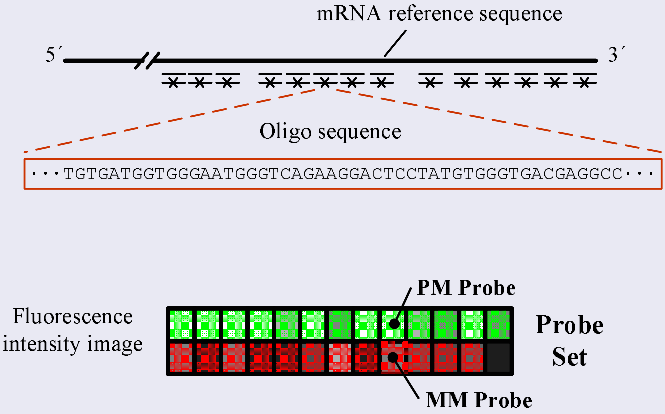

mRNA reference sequence
5′
3′
Oligo sequence
· · · TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC · · ·
PM Probe
Fluorescence intensity image
Probe Set
MM Probe

Figure: Affymetrix array schematic

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

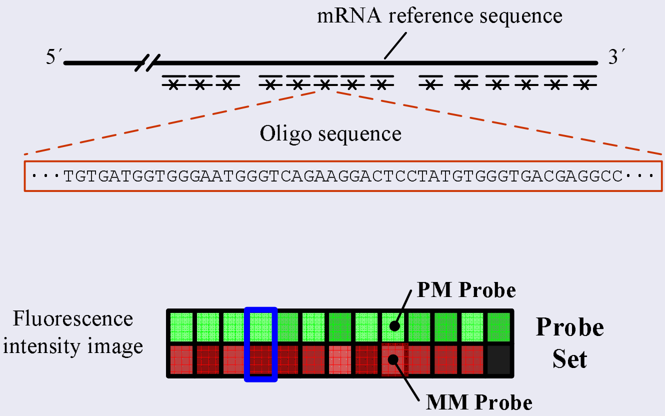

mRNA reference sequence
5′
3′
Oligo sequence
···TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
PM Probe
Fluorescence intensity image
Probe Set
MM Probe

Figure: Affymetrix array schematic

Affymetrix Arrays

Photolithography and Combinatorial Chemistry

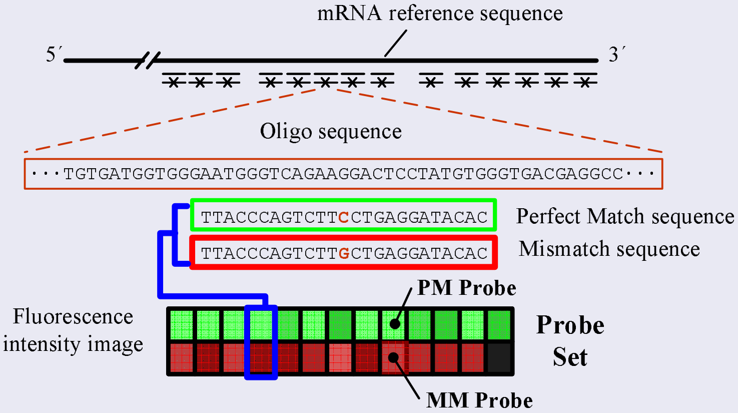

mRNA reference sequence
5′
3′
Oligo sequence
· · · TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC · · ·
TTACCCAGTCTTCCTGAGGATACAC
Perfect Match sequence
TTACCCAGTCTTGCTGAGGATACAC
Mismatch sequence
PM Probe
Fluorescence intensity image
Probe Set
MM Probe

Figure: Affymetrix array schematic

gMOS Family of Methods

Gamma Model of Signal [Milo et al., 2003, Liu et al., 2005]

- Most methods return a single expression level estimate.
- The gMOS family of methods additionally provide confidence intervals.
- This confidence intervals can the be propagated through higher level analysis.

gMOS Family of Methods II

Gamma Model of Signal

$$s_j \sim \text{Ga}\left(s_j|\alpha, b\right)$$

$$m_j \sim \text{Ga}\left(m_j|a, b\right)$$

$$y_j = m_j + s_j$$

$$y_j \sim \text{Ga}\left(y_j|a+\alpha, b\right)$$

$$\text{Ga}\left(x|a, b\right) = \frac{b^a}{\Gamma\left(a\right)} x^a \exp\left(-bx\right)$$

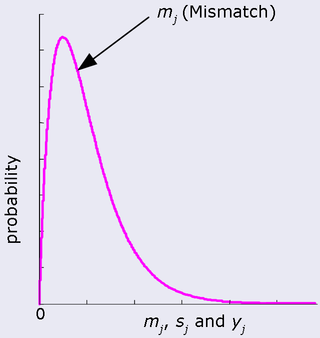


Figure: PDF of m_j, s_j and the implied distribution for y_j.

gMOS Family of Methods II

Gamma Model of Signal

$$s_j \sim \mathrm{Ga}\left(s_j | \alpha, b\right)$$

$$m_j \sim \mathrm{Ga}\left(m_j | a, b\right)$$

$$y_j = m_j + s_j$$

$$y_j \sim \mathrm{Ga}\left(y_j | a + \alpha, b\right)$$

$$\mathrm{Ga}\left(x | a, b\right) = \frac{b^a}{\Gamma\left(a\right)} x^a \exp\left(-bx\right)$$

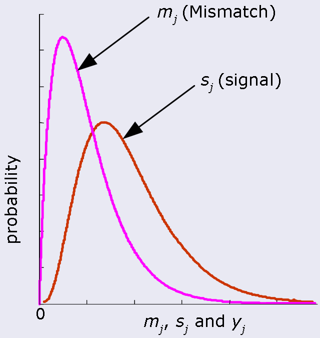


Figure: PDF of m_j, s_j and the implied distribution for y_j.

gMOS Family of Methods II

Gamma Model of Signal

$$s_j \sim \mathrm{Ga}\left(s_j | \alpha, b\right)$$

$$m_j \sim \mathrm{Ga}\left(m_j | a, b\right)$$

$$y_j = m_j + s_j$$

$$y_j \sim \mathrm{Ga}\left(y_j | a + \alpha, b\right)$$

$$\mathrm{Ga}\left(x | a, b\right) = \frac{b^a}{\Gamma\left(a\right)} x^a \exp\left(-bx\right)$$

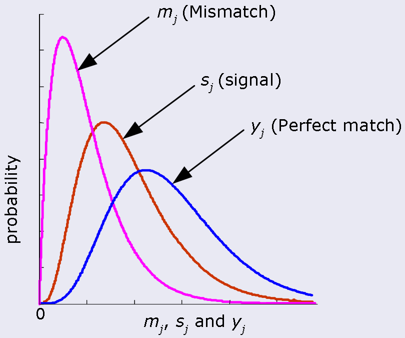


Figure: PDF of m_j, s_j and the implied distribution for y_j.

gMOS

Inferring the Signal

- Maximise likelihood with respect to α, a and b.
  - Assume independence between y_j and m_j,

$$p\left(y_j, m_j\right) = \text{Ga}\left(y_j|\alpha, b\right)\text{Ga}\left(m_j|a, b\right).$$

- Use resulting $\hat{\alpha}$ and $\hat{b}$ to give distribution over s_j.

$$p\left(s_j\right) = \text{Ga}\left(s_j|\hat{\alpha}, \hat{b}\right).$$

Modelling Probe Pair Affinity

mgMOS

- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale b_j.
- Assume $b_j \sim \mathrm{Ga}\left(b_j | c, d\right)$ and marginalise.

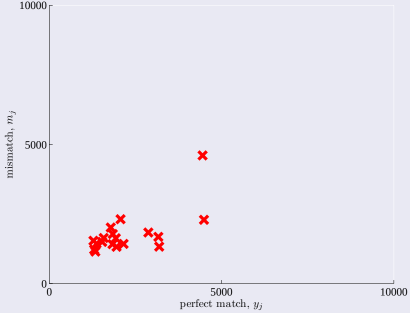


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

mgMOS

- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale b_j.
- Assume $b_j \sim \mathrm{Ga}\left(b_j|c, d\right)$ and marginalise.

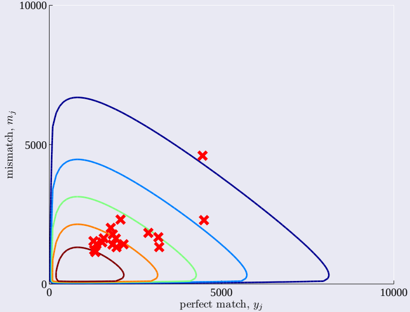


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

mgMOS

- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale b_j.
- Assume $b_j \sim \text{Ga}\left(b_j | c, d\right)$ and marginalise.

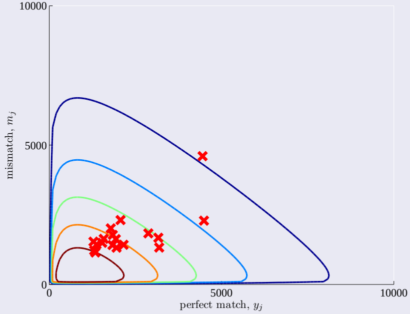


Figure: Correlation of PM (y_j) and MM (m_j).

Modelling Probe Pair Affinity

mgMOS

- y_j and m_j are correlated.
- gMOS makes an independence assumption.
- Correlations arise through shared binding affinity (scale).
- Assume each probe pair has a shared scale b_j.
- Assume $b_j \sim \text{Ga}\left(b_j|c, d\right)$ and marginalise.

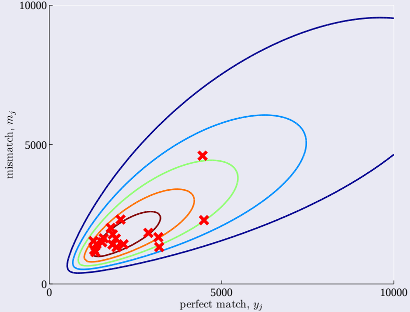


Figure: Correlation of PM (y_j) and MM (m_j).

Specific Binding to Mismatch

Mismatch Effected by Signal

- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

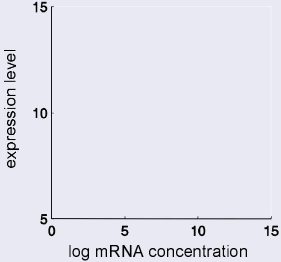


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

Specific Binding to Mismatch

Mismatch Effected by Signal

- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

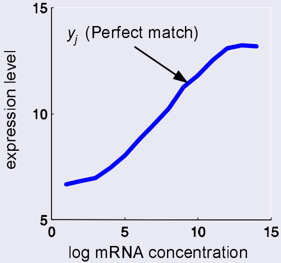


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

Specific Binding to Mismatch

Mismatch Effected by Signal

- Affymetrix Latin Square Spike-In data set.
- The perfect match responds to increasing mRNA.
- But so does the mismatch.

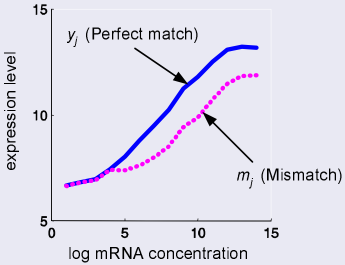


Figure: The perfect match goes up with the mRNA concentration as expected. But so does the mismatch.

Specific Binding and Multiple Arrays

multi-mgMOS

- Specific Binding to MM probe:
  - Introduce parameter ϕ and assume

$$y_j \sim \text{Ga}\left(y_j|a+\alpha, b_j\right), \quad m_j \sim \text{Ga}\left(m_j|a+\phi\alpha, b_j\right)$$

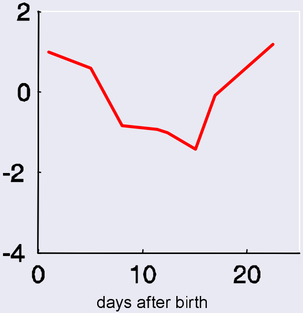
  - Log normal prior for ϕ and seek a MAP solution.
- Multiple arrays:
  - Still take $b_j \sim \text{Ga}\left(b_j|c, d\right)$ but **share c and d parameters across chips.**

Mouse Data Set

http://www.ncbi.nlm.nih.gov/projects/geo

Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

	Root Mean Square Error	
RMSE	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**



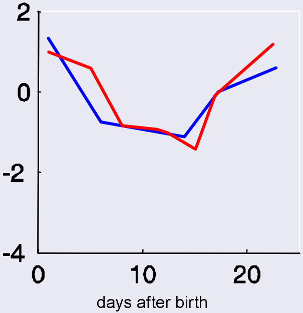
Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

http://www.ncbi.nlm.nih.gov/projects/geo

Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**



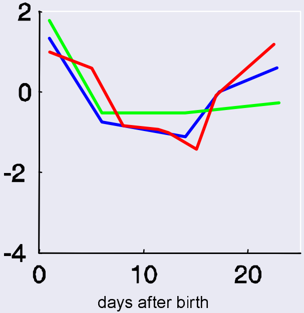
Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

http://www.ncbi.nlm.nih.gov/projects/geo

Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error	
	qr-PCR	x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	0.601	0.233



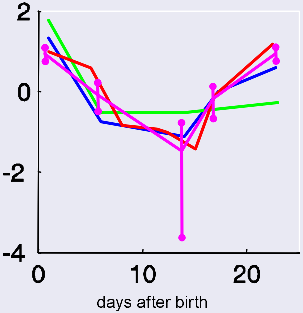
Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Mouse Data Set

http://www.ncbi.nlm.nih.gov/projects/geo

Mouse back skin mRNA expression profile for Dab2 [Lin et al., 2004].

RMSE	Root Mean Square Error qr-PCR	 x-probe set
MAS 5.0	0.656	0.360
GCRMA	0.694	0.370
multi-mgMOS	**0.601**	**0.233**



Prediction of Dab2 Expression level from qr-PCR, MAS 5.0, GCRMA and multi-mgMOS.

Differential Gene Expression

Probability of Positive Log Ratio[Liu et al., 2006]

- Differential gene expression is normally assessed with log ratios of gene expression.

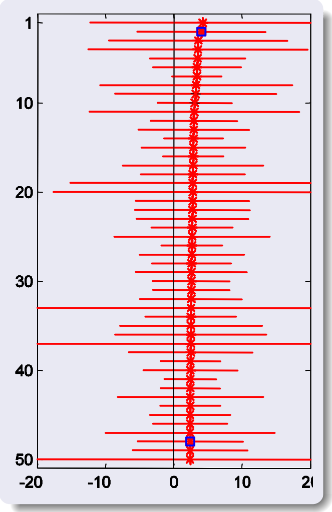
$$r_{ij} = \log \frac{s_i}{s_j}$$

- This measure is very sensitive to noise at low expresion levels.
- Use variance of expression to obtain Probability of Positive Log Ratio (PPLR).

PPLR Results

Golden spike-in dataset [Choe et al., 2005]

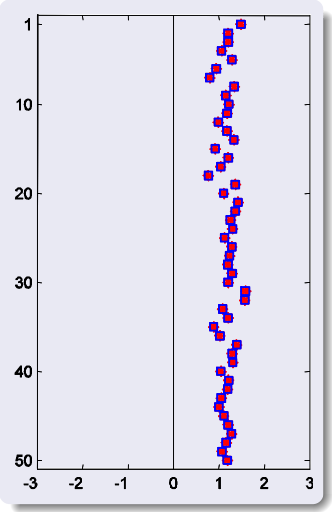
- Ranking (*y*-axis) against log ratio (*x*-axis) for.
  - **Ranking by Expected Log Ratio.**
  - **Ranking by PPLR.**
- Red stars indicate expected log ratio.
- Red lines indicate error bars.
- Blue squares indicates genes that were spiked-in.



PPLR Results

Golden spike-in dataset [Choe et al., 2005]

- Ranking (*y*-axis) against log ratio (*x*-axis) for.
  - Ranking by Expected Log Ratio.
  - **Ranking by PPLR**.
- Red stars indicate expected log ratio.
- Red lines indicate error bars.
- Blue squares indicates genes that were spiked-in.



Cleaning up Profiles

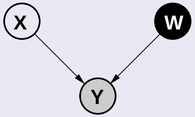
Converting Noisy Profiles to Clean

- If we can 'clean up' the profiles we can use in other methods.
- Construct a probabilistic model for the data and corruption process.
- Work with posterior distribution over cleaned up profile.
- We designed a heteroschedastic Probabilistic PCA for doing this [Sanguinetti et al., 2005].

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, **X**.
  - Integrate out *latent variables*.



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

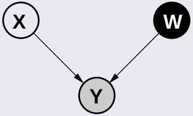
$$p\left(\mathbf{X}\right)=\prod_{i=1}^{n}N\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p\left(\mathbf{Y}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\top}+\sigma^{2}\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

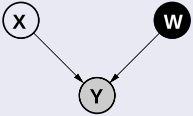
$$p\left(\mathbf{X}\right)=\prod_{i=1}^{n}N\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p\left(\mathbf{Y}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\top}+\sigma^{2}\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, **X**.
  - Integrate out *latent variables*.



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

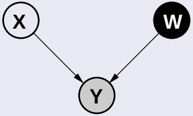
$$p\left(\mathbf{X}\right)=\prod_{i=1}^{n}N\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p\left(\mathbf{Y}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\top}+\sigma^{2}\mathbf{I}\right)$$

Probabilistic PCA

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- Latent variable approach:
  - Define Gaussian prior over *latent space*, $\mathbf{X}$.
  - Integrate out *latent variables*.



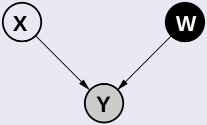
$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

$$p\left(\mathbf{X}\right) = \prod_{i=1}^{n} N\left(\mathbf{x}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Probabilistic PCA II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]



$$p\left(\mathbf{Y}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\top}+\sigma^{2}\mathbf{I}\right)$$

Probabilistic PCA II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log\left|\mathbf{C}\right| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\tilde{\mathbf{Y}}^{\mathsf{T}}\tilde{\mathbf{Y}}\right) + \mathrm{const.}$$

Where $\tilde{\mathbf{Y}}$ is the matrix $\mathbf{Y}$ with $\boldsymbol{\mu}$removed. If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\tilde{\mathbf{Y}}^{\mathsf{T}}\tilde{\mathbf{Y}}$ and the corresponding eigenvalues are Λ_q,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{V}^{\mathsf{T}}, \quad \mathbf{L} = \left(\Lambda_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

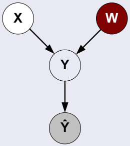
where $\mathbf{V}$ is an arbitrary rotation matrix.

$$\boldsymbol{\mu} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_{i,:}$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
  - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables.*



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}\right)$$

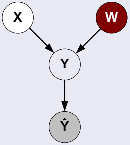
$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right) = N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:},\mathbf{D}_{i}\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\mathrm{T}}+\sigma^{2}\mathbf{I}+\mathbf{D}_{i}\right)$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
  - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables.*



$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}+\boldsymbol{\mu},\sigma^2\mathbf{I}\right)$$

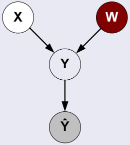
$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right) = N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:},\mathbf{D}_i\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\mathrm{T}}+\sigma^2\mathbf{I}+\mathbf{D}_i\right)$$

Heteroschedastic Probabilistic PCA

Heteroschedastic PPCA

- Define *linear-Gaussian relationship* between latent variables and $\mathbf{Y}$.
- Define a *further* Gaussian relationship to corrupted profiles $\hat{\mathbf{Y}}$.
    - $\mathbf{D}_i$ is a diagonal matrix of estimated variances.
- Integrate out *latent variables*.



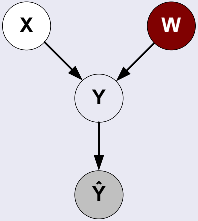
$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$$

$$p\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}\right) = N\left(\hat{\mathbf{y}}_{i,:}|\mathbf{y}_{i,:}, \mathbf{D}_i\right)$$

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I} + \mathbf{D}_i\right)$$

Heteroschedastic PPCA II

Heteroschedastic PPCA Max. Likelihood Soln [Sanguinetti et al., 2005]



$$p\left(\mathbf{Y}|\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu},\mathbf{W}\mathbf{W}^{\top}+\sigma^{2}\mathbf{I}+\mathbf{D}_{i}\right)$$

Heteroschedastic PPCA II

Heteroschedastic PPCA Max. Likelihood Soln [Sanguinetti et al., 2005]

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I} + \mathbf{D}_i\right)$$

- Can no longer solve via eigenvalue problem.
- We use an EM algorithm.
  - A major problem is the strong correlation between **W** and $\boldsymbol{\mu}$.
  - We use some tricks to speed up convergence.
- Software available in R and MATLAB.

Heteroschedastic PPCA Results

Mouse Cochlear Dataset

- Data from a conditionally imortal cell lin extracted from mouse cochlear epitheleal cells.
- Twelve samples from 14 days of differentiation after extration at E13.5 [Rivolta et al., 2002].
- Experimental setup:
  - Perform HPPCA/PCA on the data.
  - Extract 50 genes most associated with 2nd principal component
  - Cluster original profiles and reconstructed profiles.

Heteroschedastic PPCA Results

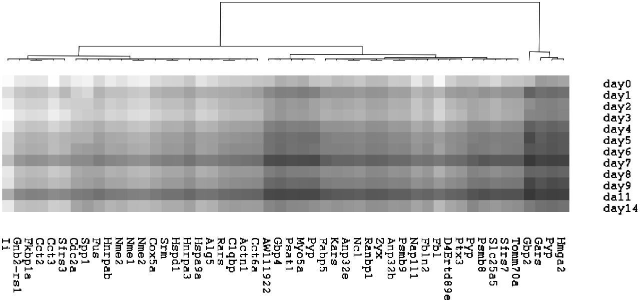


Figure: Hierarchical Clustering on Corrected Profiles.

Heteroschedastic PPCA Results

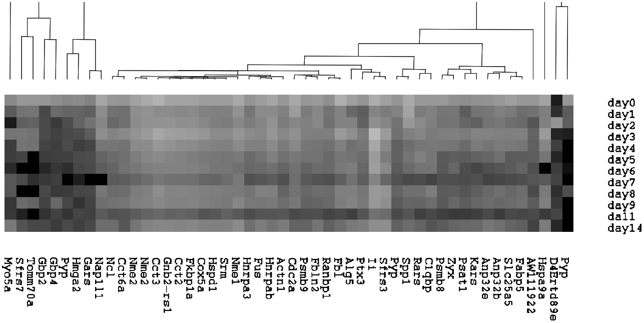


Figure: Hierarchical Clustering on Uncorrected Profiles.

Heteroschedastic PPCA Results

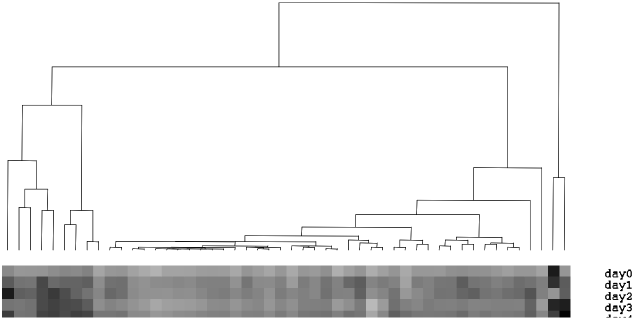


Figure: Hierarchical Clustering on Uncorrected Profiles.

Heteroschedastic PPCA Results

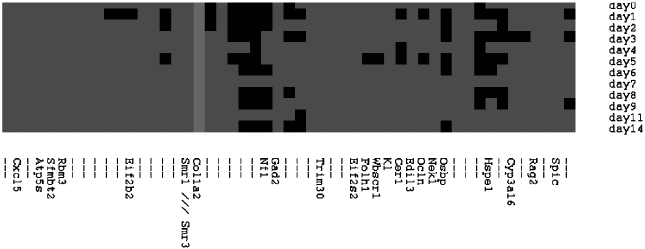


Figure: Hierarchical Clustering on genes selected by normal PCA.

Heteroschedastic PPCA Results

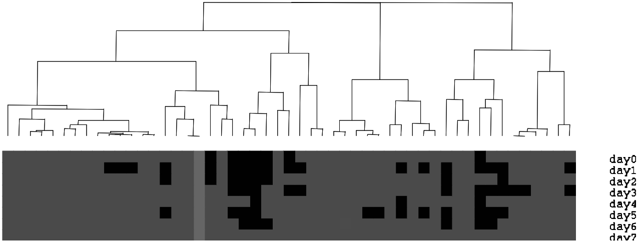


Figure: Hierarchical Clustering on genes selected by normal PCA.

Transcription Factor Activities

Inferring Activitiy of Transcription Factors

- Transcription factors control the expression of genes.
- Knowledge of their ‘activity’ is key to understanding the mechanism behind biological processes.
- Transcription factors are proteins — activity is a combination of their concentration and effect.
- The mRNA concentration of a given transcription factor may be known but:
  - Transcription factors are often lowly expressed — mRNA concentrations difficult to measure.
  - Transcription factors are often post-transcriptionally regulated.

ChIP Microarrays

Chromatine Immunoprecipitation (ChIP) Microarrays

- ChIP Microarrays tell us which TFs bind to which genes under certain conditions.
- In effect this gives a structure for the regulatory network.
- Combine this information with gene expression data to obtain transcription factor activities (TFA).

Transcription Factor Activities

Evaluating Activities of Transcription Factors

- Several approaches based on regression [Liao et al., 2003, Gao et al., 2004, Boulesteix and Strimmer, 2005, Alter and Golub, 2004]
- Assume a gene's expresion is given by a linear relationship

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \epsilon_i.$$

$$\mathbf{y}_i \in \Re^{T\times 1} \text{ is the expression profile of the } i\text{th gene,}$$

$$\mathbf{x}_i \in \{0,1\}^{q\times 1} \text{ indicates which transctiption factors bind to the } i\text{th gene}$$

$$\mathbf{B} \in \Re^{T\times q} \text{ is the matrix of TFAs.}$$

$$\epsilon_i \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

- Problem: the matrix **B** is *not* gene specific. It gives average TFA across genes.

Gene Specific TFAs

Associate TFAs to Genes [Sanguinetti et al., 2006]

- Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \epsilon_i.$$

- Parameter Explosion
  - Assume prior distribution for $\mathbf{B}_i$.

$$p(\mathbf{B}) = \prod_{i=1}^{N} p(\mathbf{B}_i) = \prod_{i=1}^{N} \prod_{t=1}^{T} p(\mathbf{b}_{i,t})$$

$$p(\mathbf{b}_{i,t}) = N(\mathbf{b}_{i,t} | \mathbf{0}, \Sigma)$$

$\mathbf{b}_{i,t} \in \Re^{q \times 1}$ is the vector of TFAs for each TF associated with the ith gene at time t

Gene Specific TFAs

Associate TFAs to Genes [Sanguinetti et al., 2006]

- Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \epsilon_i.$$

- Parameter Explosion
  - Assume prior distribution for $\mathbf{B}_i$.

$$p(\mathbf{B}) = \prod_{i=1}^{N} p(\mathbf{B}_i) = \prod_{i=1}^{N} \prod_{t=1}^{T} p(\mathbf{b}_{i,t})$$

$$p(\mathbf{b}_{i,t}) = N(\mathbf{b}_{i,t} | \mathbf{0}, \Sigma)$$

$\mathbf{b}_{i,t} \in \Re^{q \times 1}$ is the vector of TFAs for each TF associated with the ith gene at time t

Gene Specific TFAs

Associate TFAs to Genes [Sanguinetti et al., 2006]

- Intoduce gene specific TFAs,

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{x}_i + \epsilon_i.$$

- Parameter Explosion
  - Assume prior distribution for $\mathbf{B}_i$.

$$p(\mathbf{B}) = \prod_{i=1}^{N} p(\mathbf{B}_i) = \prod_{i=1}^{N} \prod_{t=1}^{T} p(\mathbf{b}_{i,t})$$

$$p(\mathbf{b}_{i,t}) = N(\mathbf{b}_{i,t} | \mathbf{0}, \Sigma)$$

$\mathbf{b}_{i,t} \in \Re^{q \times 1}$ is the vector of TFAs for each TF associated with the ith gene at time t

Temporal Continuity of TFAs

Time Course Experiments

- Introduce concept of temporal conitiuity with Gaussian distribution.

$$p\left(\mathbf{b}_{i,t}|\mathbf{b}_{i,t-1}\right) = N\left(\mathbf{b}_{i,t}|\gamma\mathbf{b}_{i,t-1} + (1-\gamma)\boldsymbol{\mu}, \left(1-\gamma^2\right)\Sigma\right)$$

The temporal continuity, γ is between 0 and 1.

Temporal Continuity of TFAs II

Effect of γ

- When $\gamma = 0$ we recover

$$p\left(\mathbf{b}_{i,t}\right) = N\left(\mathbf{b}_{i,t}|\boldsymbol{\mu}, \Sigma\right)$$

  which is equivalent to the original independent model.
- As $\gamma \rightarrow 1$ we recover

$$p\left(\mathbf{b}_{i,t}|\mathbf{b}_{i,t-1}\right) = \lim_{\sigma^2 \rightarrow 0} N\left(\mathbf{b}_{i,t}|\mathbf{b}_{i,t-1}, \sigma^2\mathbf{I}\right)$$

  which is appropriate if the ‘time points’ are in fact biological replicates.

Results on TFAs

Yeast Cell Cycle Data with ChIP-on-chip 204 TFs

- Yeast cell cycle cdc15 data set [Spellman et al., 1998].
- ChIP on chip from 113 TFs [Lee et al., 2002].
- 24 experimental points in time series data.
- Compare with non-specific TFAs obtained by Regression.

Results on TFAs II

Graphs of TFAs

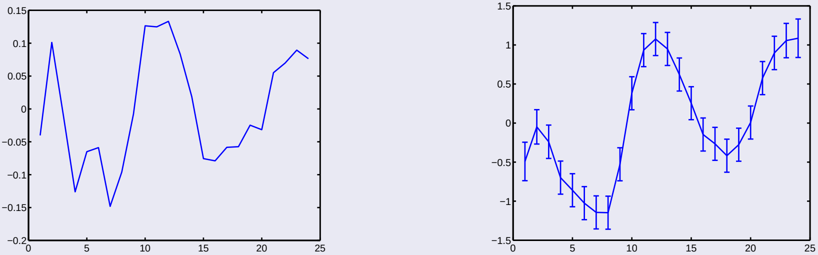


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA for averge of $\mathbf{B}_i$ across genes.

Results on TFAs II

Graphs of TFAs

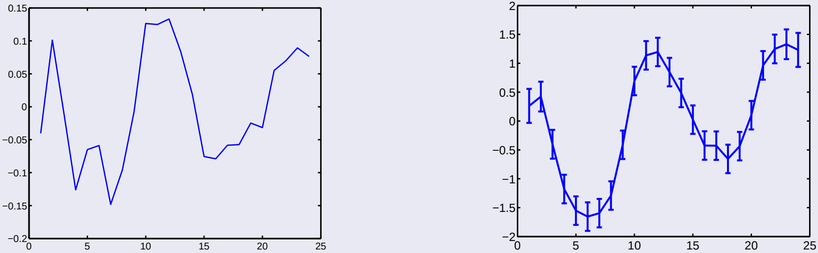


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA SCW11.

Results on TFAs II

Graphs of TFAs

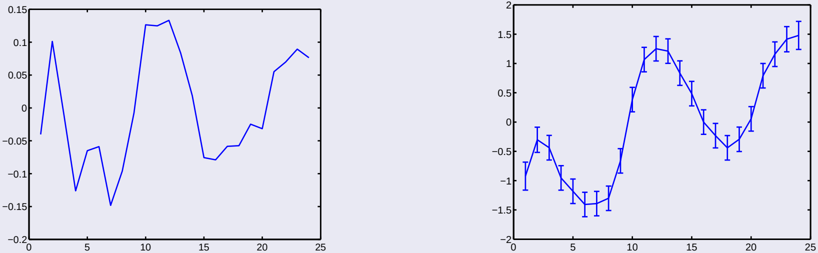


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA CTS1.

Results on TFAs II

Graphs of TFAs

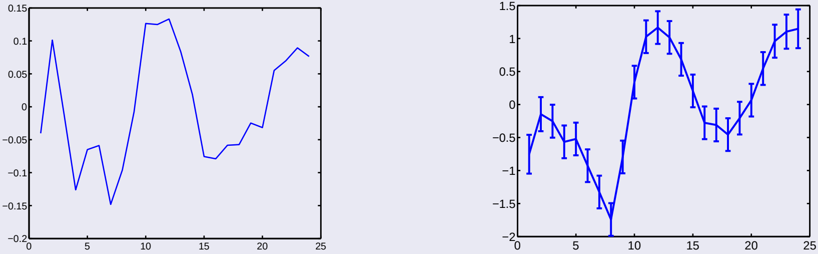


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA YER124C.

Results on TFAs II

Graphs of TFAs

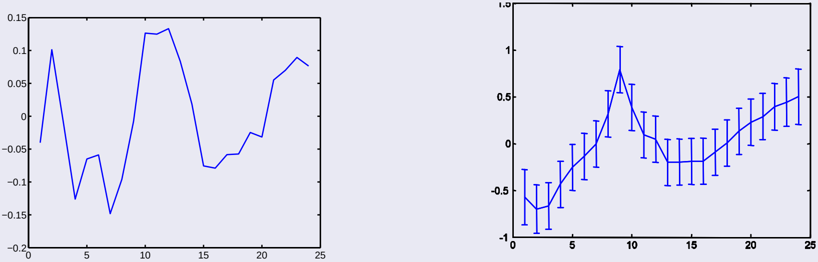


Figure: TFAs of ACE2 from the Spellman data. *Left:* TFA obtained by regression *Right:* gene specific TFA YKL51C.

Separation of Concentration and Effect

Splitting the Activity into Component Parts

- TFA is a combination of:
  - TF concentration.
  - TF effect.
- Model expression by splitting the two:

$$\mathbf{y}_i = \left(\mathbf{B} \odot \mathbf{X}\right) \mathbf{c}_t + \boldsymbol{\epsilon}_t$$

where $\odot$ is the Hadamard (element by element) product.

$\mathbf{B} \in \Re^{N\times q}$ is a matrix of each TFs effect on each gene.

$\mathbf{c}_t \in \Re^{q\times 1}$ is concentration of each TF at time t.

- Bayesian treatment of **c** and **B** through a variational approach.

TF Concentration Results

Concentration of ACE2

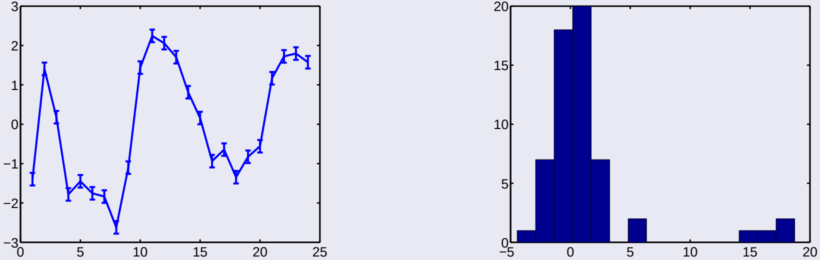


Figure: *Left:* concentration of ACE2 and *right*: effect of ACE2 on its target genes as a histogram.

TF Concentration Results II

Nice ACE2 Stories in Results

- ACE2 four most significant targets: CTS1, DSE1, DSE2, SCW11.
  - Evidence to back this up comes from CO data base.
  - CTS1 relationship is known.
  - DSE1 and DSE2 are involved in cell wall degradation causing daughter to seperate from parent.
  - SCW11's function is unclear but protein is localised at cell wall.
- Negative regulation of NCE4
  - Not documented, but ACE2 terminates mitosis & NCE4 ensures DNA stability during replication

More Complex Model

Complex Models on Small Networks

- Simple linear models allow genome wide analysis of TFAs.
- We now consider a more complex model on a much smaller network.

Differential Equation Model

Inference of p53 Concentration

- p53 is an important in cancer.
- Many targets of p53 are not shared with other TFs.
- Consider more complex model in the simple p53 network.

Differential Equation model

- Simple linear model differential equation model recently used by Barenco et al. [2006].
- They inferred transcription factor concentrations using Markov Chain Monte Carlo (10^7 iterations).
- We repeat their experiments with Gaussian processes.

Simple Linear Model

Linear model of regulation

$$\frac{dy_i\left(t\right)}{dt} = B_i + S_i f\left(t\right) - D_i y_i\left(t\right)$$

where:

$y_i\left(t\right)$ — expression of the ith gene at time t.

$f\left(t\right)$ — concentration of the transcription factor at time t.

D_i — gene's decay rate.

B_i — basal transcription rate.

S_i — sensitivity to the transcription factor.

Equation Solution

Solve via Laplace Transforms

- Solution to the equation:

$$y_i\left(t\right)=\frac{B_i}{D_i}+S_i\exp\left(-D_it\right)\int_0^t f\left(u\right)\exp\left(D_iu\right)du.$$

If $f\left(t\right)$ is a zero mean Gaussian process then $y_i\left(t\right)$ is also a Gaussian process with mean $\frac{B_i}{D_i}$.

Two Properties of GPs

Integral of Gaussian Process

The integral of a GP is also a GP,

$$f(t) \sim N(\mathbf{0}, \mathbf{K}_{ff})$$

and

$$g(t) = \int_0^t f(u)\, du$$

then

$$g(t) \sim N(\mathbf{0}, \mathbf{K}_{gg}),$$

where

$$k_{gg}(t, t') = \int_0^t \int_0^{t'} k_{ff}(u, u')\, du du'$$

Two Properties of GPs

Product with deterministic function

The integral of a GP is also a GP,

$$f(t) \sim N(\mathbf{0}, \mathbf{K}_{ff}),$$

and

$$g(t) = f(t)\, h(t)$$

where $h(t)$ is a deterministic function then,

$$g(t) \sim N(\mathbf{0}, \mathbf{K}_{gg}),$$

where

$$k_{gg}(t, t') = h(t)\, k_{ff}(t, t')\, h(t')$$

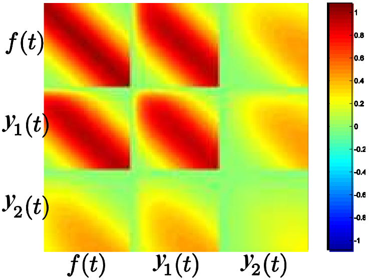
Covariance for Transcription Model

RBF Kernel function for $f(t)$

$$y_i(t) = \frac{B_i}{D_i} + S_i \exp(-D_i t) \int_0^t f(u) \exp(D_i u)\, du.$$

- Joint distribution for $x_1(t)$, $x_2(t)$ and $f(t)$.
- Here:

D_1	S_1	D_2	S_2
5	5	0.5	0.5



Joint Sampling of $y(t)$ and $f(t)$ from Covariance

gpsimTest

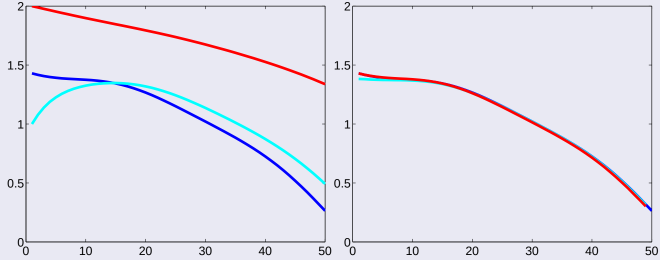


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

gpsimTest

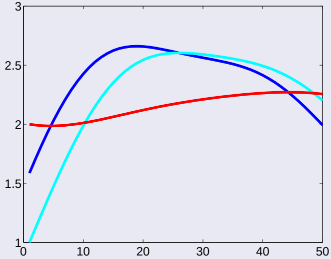


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

gpsimTest

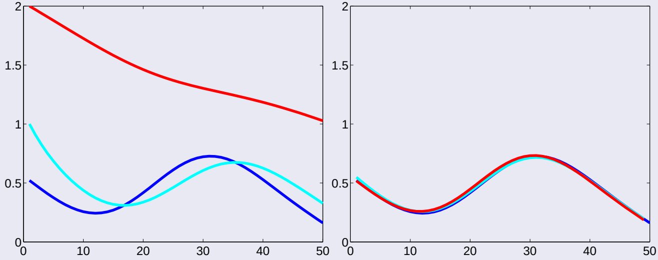


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Results — Transcription Rates

Estimation of Equation Parameters demBarenco1

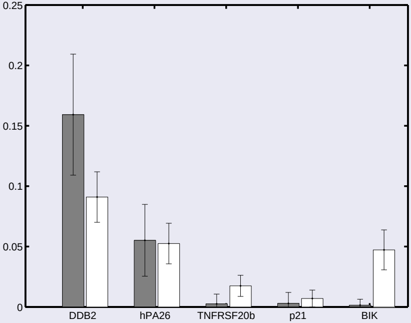


Figure: Basal transcription rates. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

Estimation of Equation Parameters `demBarenco1`

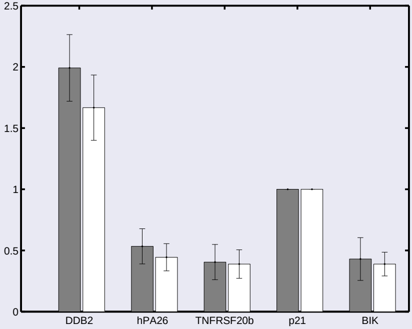


Figure: Sensitivities. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

Estimation of Equation Parameters demBarenco1

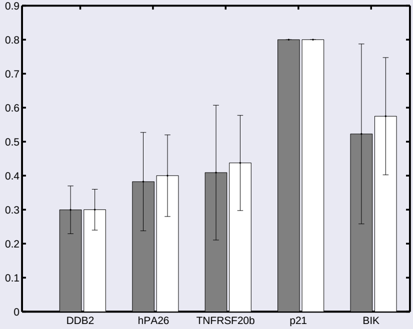
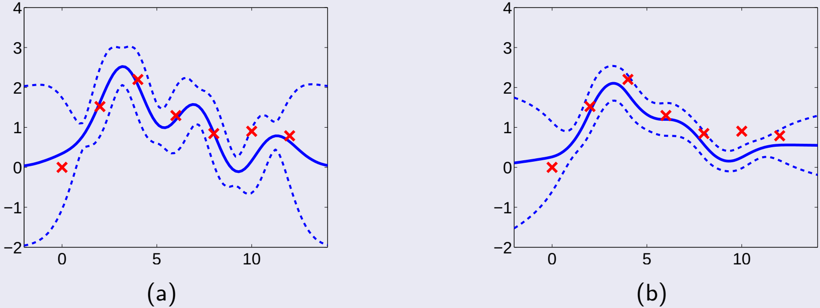


Figure: Decays. Our results (black) compared with Barenco et al. [2006] (white).

Results — Protein Concentration

Prediction with error bars of protein concentration:
$p\left(\mathbf{f}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\right)$



(a)

(b)

Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Results — Positive Constrained

GP predictions in log space.

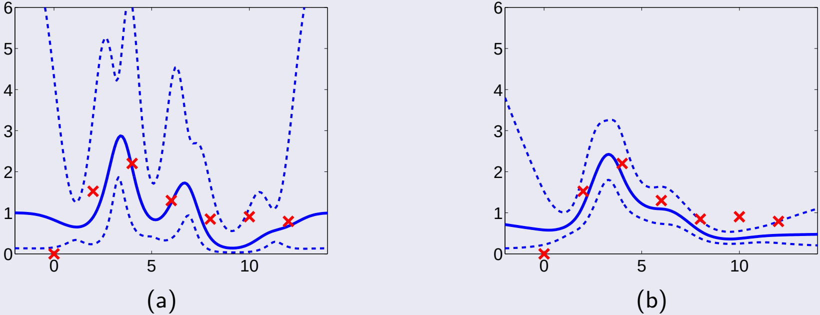


Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Transcription Model Summary

Progress so far and Future work

- Elegant solution of a problem with indirect observations.
- Already extended to non-linear response equations (using Laplace approximation).
- Expect to extend it to systems with *multiple transcription factors*.
- Gives results in 13 minutes vs 10^7 Monte-Carlo iterations.

Summary

PUMA: Propagation of Uncertainty in Microarray Analysis

- Level of Noise in the Array can be Assesed (gMOS methods).
- Probabilistic Models can:
  - Improve selection of over-expressed genes (PPLR).
  - Clean up gene expression profiles (NPPCA).
- Simple (log-linear) probabilistic models can be used with network connectivity data to
  - To infer *genome wide* transcription factor activities (chipdyno).
  - To infer *genome wide* transcription factor protein concentrations (chipvar).
- Gaussian processes & differential equations for complex interations.
- And finally ...

Acknowledgements

Team:

- Prinicipal Investigators
  - Neil Lawrence and Magnus Rattray
- gMOS family of Methods and PPLR
  - Xuejun Liu and Marta Milo
- Uncertainty Propagation through PCA
  - Marta Milo and Guido Sanguinetti
- Inference of Transcription Factor Activities
  - Guido Sanguinetti

References

O. Alter and G. H. Golub. Integrative analysis of genome-scale data using pseudoinverse projection predicts novel correlation between dna replication and rna transcription. *Proceedings of the National Academy of Sciences USA*, 101(47):16577–16582, 2004.

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

A.-L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model.*, 2(23): 1471–16582, 2005.

S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(R16), 2005.

F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5(31):

First Frame