Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components

Thomas Tüchler\textsuperscript{1}  Florian Klinglmüller\textsuperscript{2}  Peter Sykacek\textsuperscript{1}  David P. Kreil\textsuperscript{1}

\textsuperscript{1} WWTF Chair of Bioinformatics, BOKU University, 1190 Vienna, Austria.
\textsuperscript{2} Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

5 December 2008
Evaluation approaches

- Spike-based approaches
  Knowing concentration of particular RNA species

- Non spike-based approaches
  Knowing features about whole samples
Evaluation approaches

- Spike-based approaches
  Knowing concentration of particular RNA species
- Non spike-based approaches
  Knowing features about whole samples

→ Ability to detect subtle biological differences
The EMERALD dataset

- 6 rats (genetically different)
- 4 titrations of two tissues (liver to kidney 4:0, 3:1, 1:3, 0:4)
- 3 technical replicates per sample
- 3 Platforms (Affymetrix, Agilent, Illumina)
- 216 microarrays

→ Biological vs. technical variance
Preprocessing
  - Comparable measurements
  - Clean data

Modelling
  - Exploiting mixture design

Investigating
  - Biological vs. technical variance
  - Impact of signal intensity and Normalization
Exploring the data

All platforms are affected by outlier slides

▶ Affymetrix: 3B-3
▶ Agilent: third replicate series (hybridization names ‘...-3’) conducted by operator ‘A’ (high ‘AmpLabelingInputMass’)
▶ Illumina: low cRNA yield (3 of 6), plate location 3 (4 of 6) and hybridization date 10/04/08 (5 of 6)
Modelling the mixture factor

Mixing total RNA

\[
C_{\text{mRNA}} = 0.75 \cdot a \cdot A_{\text{totRNA}} + 0.25 \cdot b \cdot B_{\text{totRNA}}
\]

\[
\rho = b/a
\]

(from Shippy, 2006)
Modelling the mixture factor

Mixing total RNA

Measuring mRNA

\[
C_{\text{mRNA}} = 0.75 \cdot a \cdot A_{\text{totRNA}} + 0.25 \cdot b \cdot B_{\text{totRNA}}
\]

\[
\rho = \frac{b}{a}
\]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>100</td>
<td>82</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>Kidney</td>
<td>0</td>
<td>18</td>
<td>67</td>
<td>100</td>
</tr>
</tbody>
</table>

\[\rho = \frac{2}{3}\]

→ Titration ≠ Mixture ratio

(from Shippy, 2006)
Modelling the EMERALD data

\[
\begin{align*}
\mu_{\text{Lig}} & \sim \mathcal{N}(\mu_{\text{Lg}}, \lambda_{\text{Lg}}^{-1}) \\
\mu_{\text{Kig}} & \sim \mathcal{N}(\mu_{\text{Kg}}, \lambda_{\text{Kg}}^{-1}) \\
\mathbf{x}_{\text{Lig}} & \sim \mathcal{N}(\mu_{\text{Lig}}, \lambda_{\text{t}}^{-1}) \\
\mathbf{x}_{\text{Kig}} & \sim \mathcal{N}(\mu_{\text{Kig}}, \lambda_{\text{t}}^{-1}) \\
\mathbf{m}_{\text{ig}} & = f(\mu_{\text{Lig}}, \mu_{\text{Kig}}, \rho) \\
\mathbf{x}_{\text{M1ig}} & \sim \mathcal{N}(\mathbf{m}_{1\text{ig}}, \lambda_{\text{t}}^{-1}) \\
\mathbf{x}_{\text{M2ig}} & \sim \mathcal{N}(\mathbf{m}_{2\text{ig}}, \lambda_{\text{t}}^{-1})
\end{align*}
\]

Directed Acyclic Graph (DAG)
Inference using Variational Bayes

Approximate true posterior $p(\theta|x)$ by factorized $Q(\theta)$:

$$Q(\theta) = \prod_i Q_i(\theta_i)$$

Optimizing free energy between $p(\theta|x)$ and $Q(\theta)$:

$$FE = -\int Q(\theta) \ln \frac{Q(\theta)}{p(x,\theta)} \, d\theta$$

Updates for $\mu$ and $\sigma$ of a Gaussian (from MacKay, 2003)
Validating the implementation

Pros
► Full posterior distributions
► Fast convergence
► Monitor convergence
► Model comparison yard stick

Cons
► Approximation

Simulation validating retrieval and conv.
Results within platform and intensity

- Biological variance detectable
- Biological < technical variance
- More mRNA in kidney samples; $\rho > 1$ (cf. Liggett, 2008)

\[
0 < \frac{\text{var}_{\text{bio}}}{\text{var}_{\text{bio}} + \text{var}_{\text{tec}}} < 1
\]
Results across platforms and intensities

Baseline

Percentage of genes with biological > technical variance

- Platform and intensity dependent
- Normalization dependent
Results across platforms and intensities

Percentage of genes with biological > technical variance

- Platform and intensity dependent
- Normalization dependent
Results across platforms and intensities

Percentage of genes with biological > technical variance

- Platform and intensity dependent
- Normalization dependent
Conclusions and Outlook

Methods

▶ Detecting biological differences as performance measure
▶ VB model exploiting mixture design
Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design
  → Efficient Bayesian inference for genome size data
Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design
  → Efficient Bayesian inference for genome size data

EMERALD dataset

- Biological variance detectable
- Platform and intensity dependence
Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design
  → Efficient Bayesian inference for genome size data

EMERALD dataset

- Biological variance detectable
- Platform and intensity dependence