IntegratedJM - an R package to Jointly Model the Gene-Expression and Bioassay Data, Taking Care of the Fingerprint Feature effect

Rudradev Sengupta, Nolen Joy Perualila, Ziv Shkedey

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

UseR 2017, Brussels, Belgium

07.07.2017
Research Team

University of Hasselt:
- Rudradev Sengupta
- Nolen Joy Perualila
- Theophile Bigirumurame
- Ziv Shkedy

Durham University:
- Adetayo Kasim

JnJ:
- Luc Bijnens
- Willem Talloen
- Bie Verbist
- Hinrich Göhlmann

QSTAR Consortium
Introduction to the QSTAR project.
Model Formulation.
Brief Overview

- Introduction to the QSTAR project.
- Model Formulation.
- IntegratedJM R package.
- ShinyR web application for the IntegratedJM R package.
Modern drug discovery processes involve multiple sources of high-dimensional data -> challenge of data integration.
Modern drug discovery processes involve multiple sources of high-dimensional data -> challenge of data integration.

Example: Launching a new drug in the market
- initial safety check by the pharmaceutical companies prior to later and expensive phases of drug development cycles.
- required to analyse chemical structure (fingerprint features) of the compounds, phenotypic bioactivity data for targets of interest together with the gene expression data.
Modern drug discovery processes involve multiple sources of high-dimensional data -> challenge of data integration.

Example: Launching a new drug in the market
- initial safety check by the pharmaceutical companies prior to later and expensive phases of drug development cycles.
- required to analyse chemical structure (fingerprint features) of the compounds, phenotypic bioactivity data for targets of interest together with the gene expression data.

Modelling approach (Perualila-Tan et al., 2016):
- for a specific compound, uncovers the association between gene expression and biological activity, taking into account the influence of the chemical structure of the compound on both the variables.
- allows to detect genes that are associated with the bioactivity data, facilitating the identification of potential genomic biomarkers.
How to better understand the mechanism of new compound(s), taking into account different sources of information about the compound(s) ?

The main aim was to model the association between gene expression and bioactivity variables, given a chemical structure.
Biomarker Setting

Feature-specific and gene-by-gene analysis to use gene expression in order to understand the biological processes related to the new compound.

\[ Z_{KxI} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1I} \\ z_{21} & z_{22} & \cdots & z_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ z_{K1} & z_{K2} & \cdots & z_{KI} \end{bmatrix} \]

\[ X_{JxI} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1I} \\ x_{21} & x_{22} & \cdots & x_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ x_{J1} & x_{J2} & \cdots & x_{JI} \end{bmatrix} \]

\[ Y_{BxI} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1I} \\ y_{21} & y_{22} & \cdots & y_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ y_{B1} & y_{B2} & \cdots & y_{BI} \end{bmatrix} \]

Rudradev Sengupta

IntegratedJM - an R package to Jointly Model the Gene-Expression and Bioassay Data, Taking Care of the Fingerprint Feature effect
Model Formulation

Model for the $k^{th}$ fingerprint feature, $j^{th}$ gene and $i^{th}$ compound:

\[
\begin{pmatrix}
X_{ij} \\
Y_i
\end{pmatrix}
\sim N
\left(\begin{pmatrix}
\mu_{jk} + \alpha_{jk} \times Z_{ki} \\
\mu_{yk} + \beta_k \times Z_k
\end{pmatrix}, \Sigma_{xjk} = \begin{pmatrix}
\sigma_{xk}^2 & \sigma_{XYk} \\
\sigma_{XYk} & \sigma_{yk}^2
\end{pmatrix}\right)
\]

$\alpha_{jk}$ FF Effect: the effect of the $k^{th}$ fingerprint feature on the $j^{th}$ gene.

$\beta_k$ FF Effect: the effect of the $k^{th}$ fingerprint feature on the bioassay read-out.

$\rho_{jk}$ Adjusted Association: the correlation between the $j^{th}$ gene expression and the response after adjustment for the $k^{th}$ fingerprint effects.

jmRes <- fitJM(genematrix/expressionset, activityvector, fpvector, methodMultTest = 'fdr')
Unadjusted and Adjusted Association

Given the group, the data are **not correlated**.

The correlation between response and gene is because of the grouping.

Given the groups, variables are **correlated**.

Within each level of the grouping factor the data are correlated.
The expression level versus the bioactivity for gene FGFBP1 and the corresponding plot for the residuals.

Residual plot can be turned off by setting resPlot = FALSE in the argument of the `plot1gene` function.

\[\text{p1 } \leftarrow \text{plot1gene}(\text{geneName, fpvector, fpName, responseVector, genedata, resPlot=TRUE, colP = "blue", colA = "white")}\]
The argument `Effect` implies that a subset of differentially expressed genes will be selected. The ranking argument specifies how to rank the genes within the selected subset. For the EGFR project the gene `KRAS` has the highest absolute value for unadjusted association.

```
topkGenes(jointModelResult=jmRes, subset_type ="Effect",ranking = "Pearson", k=5, sigLevel = 0.05)
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KRAS</td>
<td>-0.29738661</td>
<td>0.0008707120</td>
<td>0.6228461</td>
<td>0.3390398</td>
<td>0.07048685</td>
</tr>
<tr>
<td>MAP9</td>
<td>-0.13059460</td>
<td>0.0002435665</td>
<td>0.6160688</td>
<td>0.2942768</td>
<td>0.12538775</td>
</tr>
<tr>
<td>SMG1</td>
<td>-0.09732380</td>
<td>0.0019501488</td>
<td>0.6157263</td>
<td>0.3524950</td>
<td>0.05841850</td>
</tr>
<tr>
<td>PTER</td>
<td>-0.10177997</td>
<td>0.0019501488</td>
<td>0.6123932</td>
<td>0.3462307</td>
<td>0.06378699</td>
</tr>
<tr>
<td>ODZ3</td>
<td>-0.13710657</td>
<td>0.0055083660</td>
<td>0.5864808</td>
<td>0.3521713</td>
<td>0.05866836</td>
</tr>
</tbody>
</table>

The gene `FOSL1` is the top ranked gene in the subset of differentially expressed and correlated genes, when ranked based on the absolute value of the fingerprint effect on the gene expression.

```
topkGenes(jointModelResult=jmRes, subset_type ="Effect and Correlation", ranking = "CovEffect1", k=5, sigLevel = 0.05)
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FOSL1</td>
<td>1.1942768</td>
<td>0.005728185</td>
<td>-0.8396638</td>
<td>-0.7619553</td>
<td>4.590942e-06</td>
</tr>
<tr>
<td>FGFBP1</td>
<td>0.7872531</td>
<td>0.008023226</td>
<td>-0.8446045</td>
<td>-0.7813558</td>
<td>2.422426e-06</td>
</tr>
<tr>
<td>SEPP1</td>
<td>-0.6360825</td>
<td>0.008776129</td>
<td>0.8124630</td>
<td>0.7342640</td>
<td>7.168442e-06</td>
</tr>
<tr>
<td>SCGB2A1</td>
<td>-0.6139309</td>
<td>0.008017638</td>
<td>0.8288709</td>
<td>0.7565218</td>
<td>4.633984e-06</td>
</tr>
<tr>
<td>SH2B3</td>
<td>0.6106726</td>
<td>0.005601205</td>
<td>-0.7942338</td>
<td>-0.6880690</td>
<td>2.586543e-05</td>
</tr>
</tbody>
</table>
Top 5 differentially expressed genes with high adjusted correlation.

**multiplot** *(p1, p2, p3, p4, p5, cols=5)*
Other Available Functions

- **plotAsso** - plots the unadjusted association vs the adjusted association for all the genes.
- **plotEff** - plots the fingerprint effect on gene expression vs the adjusted association for all the genes.
- **volcano** - produces the volcano plot for logratios vs corresponding p-values.
A graphical user interface (GUI), to use the joint model implemented in the IntegratedJM R package, is also available and can be shared, if required.
Classification of genes together with the volcano plot.
ShinR Web Application: Top Genes

- Specify the variables for the analysis in the left panel.
- Top k genes and adjusted vs unadjusted association plot specification in the right panel.
Gene-specific analysis.
Concluding Remarks

- \texttt{fitJM} function contains sequential loop over genes to analyse all the genes -> plans to add an option to specify whether to execute the loop parallelly or not, in the next version.

- ShinyR web application for the package was also tested in Amazon EC2 instances -> scope to use it from your phone/tab just with an active internet connection.

References


Thank you for your attention...!!