Introduction to iBBiG

Aedin Culhane, Daniel Gusenleitner

October 17, 2016

1 iBBiG

Iterative Binary Bi-clustering of Gene sets (iBBiG) is a bi-clustering algorithm optimized for discovery of overlapping biclusters in sparse binary matrices of data (Gusenleitner et al. in review).

We have optimized this method for the discovery of modules in matrices of discretized p-values from gene set enrichment analysis (GSA) of hundreds of datasets. However, it could be applied to any binary (1,0) matrix, such as discretized p-values from any sources of binary data. We apply iBBiG to meta-GSA to enable integrated analysis over hundreds of gene expression datasets. By integrating data at the levels of GSA results, we avoid the need to match probes/genes across multiple datasets, making large scale data integration a tractable problem.

iBBiG scales well with the dimensions of meta-datasets and is tolerant to noise characteristic of genomic data. It outperformed other traditional clustering approaches (Hierarchical clustering, k-means) or biclustering methods (bimax, fabia, coalesce) when applied to simulated data.

2 Application to simulated dataset

To demonstrate iBBiG, we will use a simulated binary dataset of 400 rows x 400 columns (as described by Gusenleitner et al.), in which a 1 indicates a positive association (or \( p < 0.05 \)) between a gene set (row) and the results of a pairwise test between clinical covariates (column), and a 0 represents a lack of association.

To simulate random noise characteristically observed in genomic data, 10% random background noise (value of 1) was introduced into the matrix.

The matrix was seeded with seven artificial modules or bi-clusters (M1-M7; Figure 1) by assigning associations (value of 1) to its column and row pairs. To replicate the expected properties of real data, seeded modules partially overlapped in columns, in rows and in both rows and columns simultaneously. M1 gene sets overlap with most other modules with the exception of M3. M2 has overlapping pairwise tests with modules M4-7.

Artificial modules also have highly varying sizes and aspect ratios, including "wide" modules driven by a large number of pairwise tests and only a few gene sets and "tall" modules like M1 which consist of 25 pairwise tests and a large number of gene sets \( n = 250 \). This latter type of module might represent a complex, well-characterized biological process such as proliferation.

In a real data set the signal strength will vary both between and within modules. Variance between modules was simulated by imposing random noise (1 -> 0 replacement) with different signal strengths on the modules (Figure 1). Within a module, we expect to see a few strong signals (gene sets associated with all pairwise tests) and many weaker signals. Therefore within each module, a noise gradient was also applied so that the first gene sets had the greatest number of associations (Figure 1). This overlaid noise gradient ranged from 10 to 60% and varied between modules (Table 1).

To create this simulated data as described in Gusenleitner, et al. use the function makeArtifical which creates an object of class iBBiG, an extension of biclust.

> library(iBBiG)
> binMat<-makeArtifical()
<table>
<thead>
<tr>
<th>M1</th>
<th>250</th>
<th>25</th>
<th>0.4</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>75</td>
<td>175</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>M3</td>
<td>50</td>
<td>50</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>M4</td>
<td>40</td>
<td>40</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>M5</td>
<td>30</td>
<td>30</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>M6</td>
<td>20</td>
<td>20</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>M7</td>
<td>40</td>
<td>40</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Cluster sizes in new iBBiG (Biclust) data object
Number of Modules: 7
Rows 250 75 50 40 30 20 40
Columns 25 175 50 40 30 20 40

> binMat

An object of class iBBiG

Number of Clusters found: 7
First 5 Cluster scores and sizes:
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Score</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Number of Rows:</td>
<td>250</td>
<td>75</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Number of Columns:</td>
<td>25</td>
<td>175</td>
<td>50</td>
<td>40</td>
</tr>
</tbody>
</table>

> plot(binMat)
The class `BiClust` contains the number of clusters and two logical matrices which indicate whether a row or column are present in the cluster.

```r
> str(binMat)
Formal class 'iBBiG' [package "iBBiG"] with 8 slots
  ..@ Seeddata : num [1:400, 1:400] 1 1 1 1 1 1 1 1 1 1 ... 
  .. ..- attr(*, "dimnames")=List of 2 
  .. .. ..$ : chr [1:400] "sig_1" "sig_2" "sig_3" "sig_4" ... 
  .. .. ..$ : chr [1:400] "cov_1" "cov_2" "cov_3" "cov_4" ... 
  ..@ RowScorexNumber: num[0,0] 
  ..@ Clusterscores : num(0) 
  ..@ Parameters :List of 3 
  .. ..$ designMatrix: num [1:7, 1:6] 251 51 1 46 81 106 151 51 251 1 ... 
  .. ..- attr(*, "dimnames")=List of 2 
  .. .. ..$ : chr [1:7] "M1" "M2" "M3" "M4" ... 
  .. .. ..$ : chr [1:6] "startC" "startR" "endC" "endR" ... 
  .. ..$ nRow : num 400 
  .. ..$ nCol : num 400 
  ..@ RowxNumber : logi [1:400, 1:7] FALSE FALSE FALSE FALSE FALSE FALSE ... 
  ..@ NumberxCov : logi [1:7, 1:400] FALSE FALSE TRUE FALSE FALSE FALSE ... 
```
The matrix \textit{RowxNumber} is a logical matrix having a number of rows equal to that of \textit{binMat} and a number of columns equal to the number of detected clusters. \textit{NumberxCol} is reversed; this class has a row count equal to the number of clusters and a column count of the number of columns of \textit{binMat}.

To run iBBiG on this artificial binary matrix, simply call the function \texttt{iBBiG}. The function \texttt{plot} and \texttt{statClust} will provide a visual representation and statistical summary of the results of the cluster analysis.

\[
> \text{res}<- \text{iBBiG(binMat@Seeddata, nModules=10)}
\]

Module: 1 ... done
Module: 2 ... done
Module: 3 ... done
Module: 4 ... done
Module: 5 ... done
Module: 6 ... done
Module: 7 ... done
Module: 8 ... done
Module: 9 ... done
Module: 10 ... done

\[
> \text{plot(res)}
\]
If you wish to compare two iBBiG or Biclust results, for example a prediction and a gold standard (GS), the function \texttt{JIdist} will calculate the Jaccard Index distance between two Biclust or iBBiG result objects. By default, it calculates the distances between each column. Setting \texttt{margin = row} or \texttt{margin = both} will cause the function to calculate instead the JI distance between the rows, or an average of rows/columns.

By default, \texttt{RfunctionJIdist} returns a \texttt{data.frame} with 2 columns, the column \texttt{n} indicating which cluster was the best match (maximum JI) to each cluster of the second iBBiG object (GS). The column \texttt{JI} contains the Jaccard Index distance between the columns of these two clusters. If \texttt{best = FALSE}, the function will return the distance matrix instead of the best match.

\begin{verbatim}
> JIdist(res,binMat)
   n    JI
GS_1  1 1.000
GS_2  2 0.977
GS_3  5 1.000
GS_4  3 1.000
GS_5  7 0.967
GS_6  4 0.952
GS_7  6 0.725

> JIdist(res,binMat, margin="col", best=FALSE)
\end{verbatim}
> JIdist(res, binMat, margin="col")
    n   JI
   GS_1 1  1.000
   GS_2 2  0.977
   GS_3 5  1.000
   GS_4 3  1.000
   GS_5 7  0.967
   GS_6 4  0.952
   GS_7 6  0.725

> JIdist(res, binMat, margin="row")
    n   JI
   GS_1 1  0.764
   GS_2 2  0.800
   GS_3 5  0.860
   GS_4 3  0.767
   GS_5 7  0.550
   GS_6 4  0.559
   GS_7 6  0.569

> JIdist(res, binMat, margin="both")
    n   JI
   GS_1 1  1.764
   GS_2 2  1.777
   GS_3 5  1.860
   GS_4 3  1.767
To view the code of the function JIdist

```r
> showMethods(JIdist)
> getMethod(iBBiG:::JIdist, signature(clustObj = "iBBiG", GS = "iBBiG"))
> getMethod("JIdist", signature(clustObj="iBBiG", GS="iBBiG"))
```

To extract performance statistics between two iBBiG results, use `analyzeClust`, which will take a single iBBiG result object or a list of objects and compare these to a gold standard (another iBBiG or biclust object). Again results can be based on matches to the best row, column or both.

```r
> analyzeClust(res, binMat)
```

<table>
<thead>
<tr>
<th></th>
<th>Run</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GS_1</td>
<td>1</td>
<td>1</td>
<td>1,000</td>
<td>191</td>
<td>25</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_2</td>
<td>1</td>
<td>2</td>
<td>0.977</td>
<td>60</td>
<td>171</td>
<td>0.990</td>
</tr>
<tr>
<td>GS_3</td>
<td>1</td>
<td>5</td>
<td>1,000</td>
<td>43</td>
<td>50</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_4</td>
<td>1</td>
<td>3</td>
<td>1,000</td>
<td>36</td>
<td>40</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_5</td>
<td>1</td>
<td>7</td>
<td>0.967</td>
<td>32</td>
<td>29</td>
<td>0.998</td>
</tr>
<tr>
<td>GS_6</td>
<td>1</td>
<td>4</td>
<td>0.952</td>
<td>33</td>
<td>21</td>
<td>0.998</td>
</tr>
<tr>
<td>GS_7</td>
<td>1</td>
<td>6</td>
<td>0.725</td>
<td>40</td>
<td>29</td>
<td>0.972</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>col-specificity</th>
<th>col-PPV</th>
<th>col-NPV</th>
<th>row-accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS_1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.852</td>
</tr>
<tr>
<td>GS_2</td>
<td>1.000</td>
<td>1.000</td>
<td>0.983</td>
<td>0.962</td>
</tr>
<tr>
<td>GS_3</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.982</td>
</tr>
<tr>
<td>GS_4</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.975</td>
</tr>
<tr>
<td>GS_5</td>
<td>1.000</td>
<td>1.000</td>
<td>0.997</td>
<td>0.955</td>
</tr>
<tr>
<td>GS_6</td>
<td>0.997</td>
<td>0.952</td>
<td>1.000</td>
<td>0.962</td>
</tr>
<tr>
<td>GS_7</td>
<td>1.000</td>
<td>1.000</td>
<td>0.970</td>
<td>0.945</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>row-sensitivity</th>
<th>row-specificity</th>
<th>row-PPV</th>
<th>row-NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS_1</td>
<td>0.764</td>
<td>1.000</td>
<td>1.000</td>
<td>0.718</td>
</tr>
<tr>
<td>GS_2</td>
<td>0.800</td>
<td>1.000</td>
<td>1.000</td>
<td>0.956</td>
</tr>
<tr>
<td>GS_3</td>
<td>0.860</td>
<td>1.000</td>
<td>1.000</td>
<td>0.980</td>
</tr>
<tr>
<td>GS_4</td>
<td>0.825</td>
<td>0.992</td>
<td>0.917</td>
<td>0.981</td>
</tr>
<tr>
<td>GS_5</td>
<td>0.733</td>
<td>0.973</td>
<td>0.688</td>
<td>0.978</td>
</tr>
<tr>
<td>GS_6</td>
<td>0.950</td>
<td>0.963</td>
<td>0.576</td>
<td>0.997</td>
</tr>
<tr>
<td>GS_7</td>
<td>0.725</td>
<td>0.969</td>
<td>0.725</td>
<td>0.969</td>
</tr>
</tbody>
</table>

```r
> analyzeClust(res, binMat, margin="col")
```

<table>
<thead>
<tr>
<th></th>
<th>Run</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GS_1</td>
<td>1</td>
<td>1</td>
<td>1,000</td>
<td>191</td>
<td>25</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_2</td>
<td>1</td>
<td>2</td>
<td>0.977</td>
<td>60</td>
<td>171</td>
<td>0.990</td>
</tr>
<tr>
<td>GS_3</td>
<td>1</td>
<td>5</td>
<td>1,000</td>
<td>43</td>
<td>50</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_4</td>
<td>1</td>
<td>3</td>
<td>1,000</td>
<td>36</td>
<td>40</td>
<td>1.000</td>
</tr>
<tr>
<td>GS_5</td>
<td>1</td>
<td>7</td>
<td>0.967</td>
<td>32</td>
<td>29</td>
<td>0.998</td>
</tr>
<tr>
<td>GS_6</td>
<td>1</td>
<td>4</td>
<td>0.952</td>
<td>33</td>
<td>21</td>
<td>0.998</td>
</tr>
<tr>
<td>GS_7</td>
<td>1</td>
<td>6</td>
<td>0.725</td>
<td>40</td>
<td>29</td>
<td>0.972</td>
</tr>
</tbody>
</table>
Again to view the code of the function, you could:

```r
> showMethods(analyzeClust)
> getMethod("analyzeClust", signature(clustObj="iBBiG", GS="iBBiG"))
```

The structure of `iBBiG` differs from `BiClust` in that it contains ClusterScores. ClusterScores are the scores for each module. RowScoreNumber are the scores for each row in the cluster. `Seeddata` is a copy of `binMat`.

```r
> str(binMat)
```

Formal class 'iBBiG' [package "iBBiG"] with 8 slots
  ..@ Seeddata : num [1:400, 1:400] 1 1 1 1 1 1 0 0 1 1 ...  
  .. ..- attr(*, "dimnames")=List of 2  
  .. .. ..$ : chr [1:400] "sig_1" "sig_2" "sig_3" "sig_4" ...  
  .. .. ..$ : chr [1:400] "cov_1" "cov_2" "cov_3" "cov_4" ...  
  ..@ RowScoreNumber: num[0 , 0 ]  
  ..@ Clusterscores : num(0)  
  ..@ Parameters :List of 3  
  .. ..$ designMatrix: num [1:7, 1:6] 251 51 1 46 81 106 151 51 251 1 ...  
  .. .. ..- attr(*, "dimnames")=List of 2  
  .. .. .. ..$ : chr [1:7] "M1" "M2" "M3" "M4" ...  
  .. .. .. ..$ : chr [1:6] "startC" "startR" "endC" "endR" ...  
  .. ..$ nRow : num 400  
  .. ..$ nCol : num 400  
  ..@ RowxNumber : logi [1:400, 1:7] FALSE FALSE FALSE FALSE FALSE FALSE ...  
  ..@ NumberxCol : logi [1:7, 1:400] FALSE FALSE TRUE FALSE FALSE FALSE ...  
  ..@ Number : int 7  
  ..@ info : list()  

```r
> RowScoreNumber(res)[1:2,]
```

```
M 1 M 2 M 3 M 4 M 5 M 6 M 7 M 8 M 9 M 10 sig_1 0 0 0 0 0 29.12712 0 0 0 0 0 sig_2 0 0 0 0 0 33.04621 0 0 0 0 0
```

```r
> Clusterscores(res)
```
> Seeddata(res)[1:2,1:2]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cov_1</td>
<td>cov_2</td>
</tr>
<tr>
<td>sig_1</td>
<td>1  1</td>
</tr>
<tr>
<td>sig_2</td>
<td>1  0</td>
</tr>
</tbody>
</table>

There are also the slots for info and Parameters which can contain additional user-entered information about the analysis. We can subset or reorder the results like so:

> res[1:3]

An object of class iBBiG

Number of Clusters found: 3

First 3 Cluster scores and sizes:

<table>
<thead>
<tr>
<th>M 1</th>
<th>M 2</th>
<th>M 3</th>
<th>M 4</th>
<th>M 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1909.84107</td>
<td>2849.88069</td>
<td>578.31967</td>
<td>224.47956</td>
<td>736.70702</td>
</tr>
</tbody>
</table>

Number of Rows: 191.000 60.000 36.0000

Number of Columns: 25.000 171.000 40.0000

> res[c(4,2,1)]

An object of class iBBiG

Number of Clusters found: 3

First 3 Cluster scores and sizes:

<table>
<thead>
<tr>
<th>M 4</th>
<th>M 2</th>
<th>M 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>224.4796</td>
<td>2849.881</td>
<td>1909.841</td>
</tr>
</tbody>
</table>

Number of Rows: 33.0000 60.000 191.000

Number of Columns: 21.0000 171.000 25.000

> res[1, drop=FALSE]

An object of class iBBiG

There was one cluster found with Score 1909.841 and 191 Rows and 25 columns

### 3 Using biclust functions

An object from iBBiG extends the class biclust and can therefore use methods available to a biclust object. For example, there are several plot functions in BiClust

> class(res)
> par(mfrow=c(2,1))
> drawHeatmap2(res@Seeddata, res, number=4)
> biclustmember(res, res@Seeddata)
> biclustbarchart(res@Seeddata, Bicres=res)
> plotclust(res, res@Seeddata)

Statistical measures of biclustering performance including the Chia and Karuturi Function, Coherence measures and F Statistics are available within the biclust R packages.

There are function to process data, binarize or discretize data. For example, given gene expression data we can binarize or discretize the data matrix as follows and this can be input into iBBiG

> data(BicatYeast)
> BicatYeast[1:5,1:5]
> binarize(BicatYeast[1:5,1:5], threshold=0.2)
> discretize(BicatYeast[1:5,1:5])

The sub-matrices of each cluster can be extracted from the original matrix, using the function bicluster

> Modules<-bicluster(res@Seeddata, res, 1:3)
> str(Modules)

List of 3
$ Bicluster1: num [1:191, 1:25] 1 1 1 1 1 1 1 1 1 1 ...
 ..- attr(*, "dimnames")=List of 2
  ..$ : chr [1:191] "sig_51" "sig_52" "sig_53" "sig_54" ...
  ..$ : chr [1:25] "cov_251" "cov_252" "cov_253" "cov_254" ...
$ Bicluster2: num [1:60, 1:171] 0 1 1 1 1 0 0 1 1 1 ...
 ..- attr(*, "dimnames")=List of 2
  ..$ : chr [1:60] "sig_251" "sig_252" "sig_253" "sig_254" ...
  ..$ : chr [1:171] "cov_51" "cov_52" "cov_53" "cov_54" ...
$ Bicluster3: num [1:36, 1:40] 1 1 0 1 1 1 1 1 1 1 ...
 ..- attr(*, "dimnames")=List of 2
  ..$ : chr [1:36] "sig_46" "sig_47" "sig_48" "sig_49" ...
  ..$ : chr [1:40] "cov_46" "cov_47" "cov_48" "cov_49" ...

> Modules[[1]][1:3,1:4]

<table>
<thead>
<tr>
<th></th>
<th>cov_251</th>
<th>cov_252</th>
<th>cov_253</th>
<th>cov_254</th>
</tr>
</thead>
<tbody>
<tr>
<td>sig_51</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sig_52</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sig_53</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

To write results to a file use the following:

> writeBiclusterResults("Modules.txt", res, bicName="Output from iBBiG with default parameters", geneNames=rownames(res@Seeddata), arrayNames=colnames(res@Seeddata))

4 Session Info

- R version 3.3.1 (2016-06-21), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
• Base packages: base, datasets, grDevices, graphics, grid, methods, stats, utils
• Other packages: MASS 7.3-45, biclust 1.2.0, colorspace 1.2-7, iBBiG 1.18.0, lattice 0.20-34
• Loaded via a namespace (and not attached): ade4 1.7-4, flexclust 1.3-4, modeltools 0.2-21, parallel 3.3.1, stats4 3.3.1, tools 3.3.1, xtable 1.8-2

References