Interaction Based Homogeneity

Kircicegi Korkmaz, Volkan Atalay and Rengul Cetin-Atalay

April 30, 2018

1 Introduction

We need quantitative metrics to measure the quality of gene lists for the interpretation of clustering results as well as for the evaluation and comparison of different algorithms designed for several types of genomics or proteomics data such as microarrays. Interaction Based Homogeneity (IBH) measures the fitness of a gene list to an interaction network.

Given a gene list of n genes, we first form an adjacency matrix A whose rows and columns are genes in the list where $A_{ij} = 1$f genes i and j have an interaction in the network and $A_{ij} = 0$ otherwise.

The Interaction Based Homogeneity for a gene list $L = \{g_1, g_2, ..., g_n\}$ of size n is then calculated as:

$$IBH(L) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}}{n^2}$$

The *ibh* package contains easy-to-use methods to calculate Interaction Based Homogeneity in different cases. The user can use the predefined interactions which are taken from BioGRID[**?**] database or can provide his own interaction network. There are predefined interactions for 7 organisms: Arabidopsis thaliana, Caenerhabditis elegans, Drosophila melanogaster , Homo sapiens, Mus musculus, Saccharomyces cerevisae, and Schizosac-

charomyces pombe. When using predefined interactions, unique ids(systematic names), official names or Entrez ids can be used as identifers.

2 Functions

The ibh package contains 9 functions: *ibh, ibhBioGRID, ibhForMultipleGeneLists, ibhForMultipleGeneListsBioGRID, ibhClusterEval, ibhClusterEvalBioGRID, readDirectedInteractionsFromCsv, readUndirectedInteractionsFromCsv* and *findEntry*.

The *ibh* function can be used the Interaction Based Homogeneity(IBH) for a single gene list, users should provide their own interactions. The value of IBH is between 0 and 1, higher values indicate more similar gene lists. For example:

```
> library(ibh)
> data(ArabidopsisBioGRIDInteractionEntrezId)
> geneList <- list(839226, 817241, 824340, 832179, 818561, 831145,
+     838782, 826404)
> ibh(ArabidopsisBioGRIDInteractionEntrezId, geneList)

[1] 0.234375
```

In the above example, we provided or own interactions which is taken from the *simpIntLists* package.

When users want to use the predefined interactions which are taken from the BioGRID database, they should use the *ibhBioGRID* function. The function takes the organism name and identifier type as input as well as the gene list to be evaluated. For example:

```
> geneList <- list(839226, 817241, 824340, 832179, 818561, 831145,
+   838782, 826404)
> ibhBioGRID(geneList, organism = "arabidopsis", idType = "EntrezId")
```

```
[1] 0.234375
```

```
> geneList <- list("YJR151C", "YBL032W", "YAL040C", "YBL072C",
+     "YCL050C", "YCR009C")
> ibhBioGRID(geneList, organism = "yeast", idType = "UniqueId")
```

```
[1] 0.4722222
```

When the users want to evaluate more than one gene list at a time, they should use *ibhForMultipleGeneLists* and *ibhForMultipleGeneListsBioGRID* functions. For example:

```
> data(ArabidopsisBioGRIDInteractionEntrezId)
> listofGeneList <- list(list(839226, 817241, 824340, 832179, 818561,
+     831145, 838782, 826404), list(832018, 839226, 838824))
> ibhForMultipleGeneLists(ArabidopsisBioGRIDInteractionEntrezId,
+     listofGeneList)
```

```
[1] 0.2343750 0.4444444
```

```
> listofGeneList <- list(list("YJR151C", "YBL032W", "YAL040C",
+     "YBL072C", "YCL050C", "YCR009C"), list("YDR063W", "YDR074W",
+     "YDR080W", "YDR247W", "YGR183C", "YHL033C"), list("YOL068C",
+     "YOL015W", "YOL009C", "YOL004W", "YOR065W"))
> ibhForMultipleGeneListsBioGRID(listofGeneList, organism = "yeast",
+     idType = "UniqueId")
```

```
[1] 0.4722222 0.2222222 0.0400000
```

The package also include two functions for clustering evaluation: *ibhClusterEval* and *ibhClusterEvalBioGRID*. The user should first cluster the data and then provide the clustering result, names of genes clustered as input. When using BioGRID interactions two additional parameters, organism name and identifier type, is also needed. For example:

```
> require(yeastCC)
> require(stats)
> data(yeastCC)
> require(simpIntLists)
> data(YeastBioGRIDInteractionUniqueId)
> subset <- exprs(yeastCC)[1:50, ]
> d <- dist(subset, method = "euclidean")
> k <- kmeans(d, 3)
> ibhClusterEval(k$cluster, rownames(subset), YeastBioGRIDInteractionUniqueId)

[1] NaN NaN NaN

> ibhClusterEvalBioGRID(k$cluster, rownames(subset), organism = "yeast",
+     idType = "UniqueId")

[1] 0.111111111 0.023781213 0.009259259
```

In order to create interaction lists, the package also contains two functions for reading interactions from comma separated files: *readDirectedInteractionsFromCsv* and *readUndirectedInteractionsFromCsv*. Finally, the *findEntry* function provides a search through the interactions.