Bioconductor项目简介

及其在生物信息学中的应用

陈钢 chen.gang1983@gmail.com

中南大学计算机系

2008年12月13日

- 网址：http://www.bioconductor.org

- 网址：http://www.bioconductor.org
- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data.

- 网址：http://www.bioconductor.org
- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data.
- Bioconductor 2.3: 294个软件包，针对R2.8，新加入36个软件包，加入对新一代测序技术的支持。

- 软件（Software）：

- 软件（Software）：
  - LIMMA：基因芯片数据线性建模

- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理

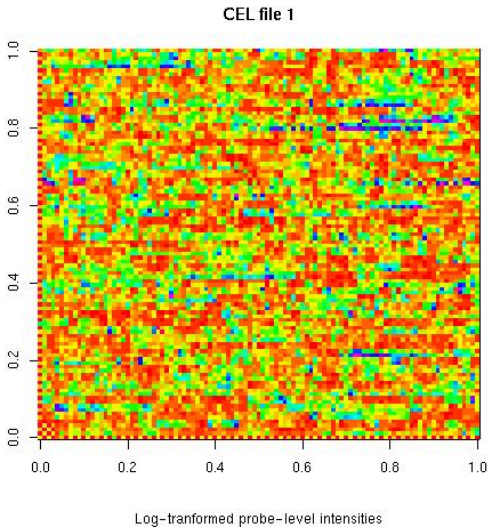
- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理
- 注释数据（AnnotationData）：

- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理
- 注释数据（AnnotationData）：
  - GO.db：Gene Ontology注释数据

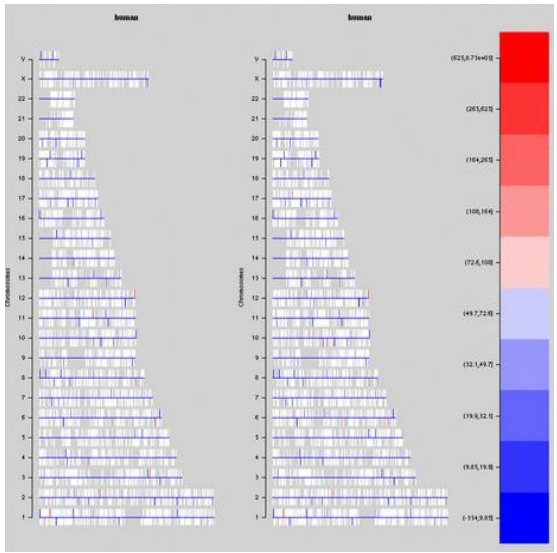
- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理
- 注释数据（AnnotationData）：
  - GO.db：Gene Ontology注释数据
  - hgu133a2.db：Affymetrix Human Genome U133A 2.0芯片的注释数据

- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理
- 注释数据（AnnotationData）：
  - GO.db：Gene Ontology注释数据
  - hgu133a2.db：Affymetrix Human Genome U133A 2.0芯片的注释数据
- 实验数据（ExperimentData）：

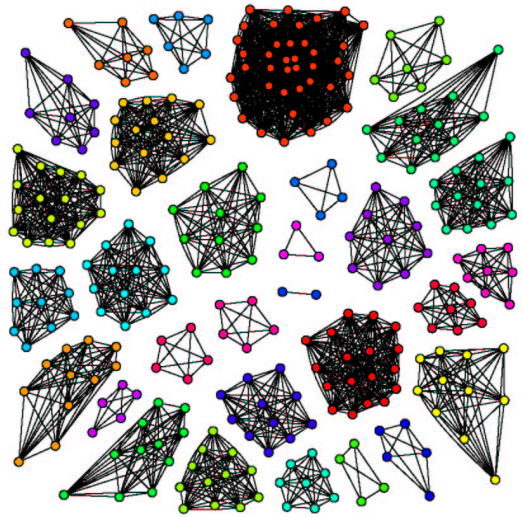
- 软件（Software）：
  - LIMMA：基因芯片数据线性建模
  - affyio：Affymetrix公司芯片数据的处理
- 注释数据（AnnotationData）：
  - GO.db：Gene Ontology注释数据
  - hgu133a2.db：Affymetrix Human Genome U133A 2.0芯片的注释数据
- 实验数据（ExperimentData）：
  - hapmapsnp5:人类单体型计划SNP数据



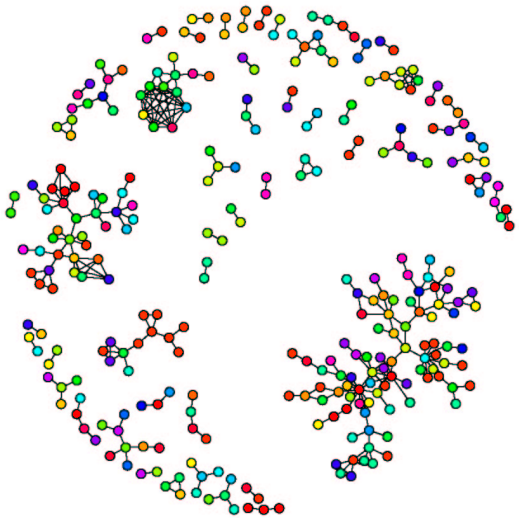
affy:Image of Affymetrix probe-level intensities, *image()* function.



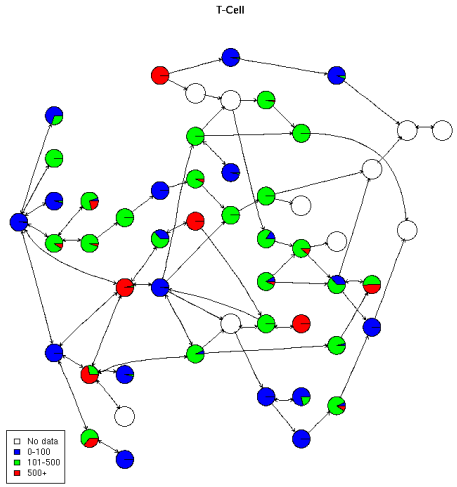
geneplotter: Comparing mean expression levels between two groups.



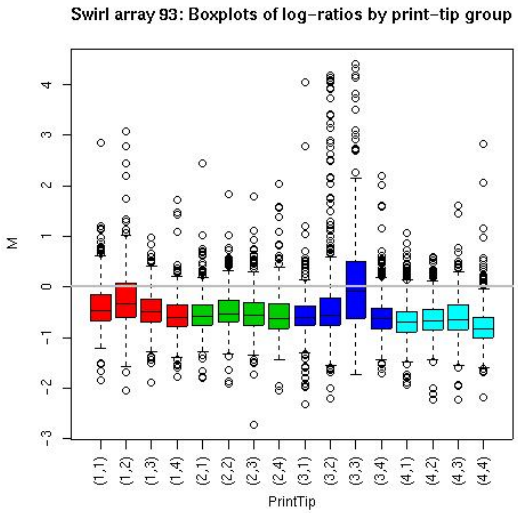
graph and **Rgraphviz**: Completely connected subgraphs for clusters of genes with similar expression profiles. Created with **graph** and plotted using **Rgraphviz**.



graph and **Rgraphviz**: Graph of interacting protein pairs with proteins (nodes) colored according to cell cycle expression profile cluster membership. Created with **graph** and plotted using **Rgraphviz**.

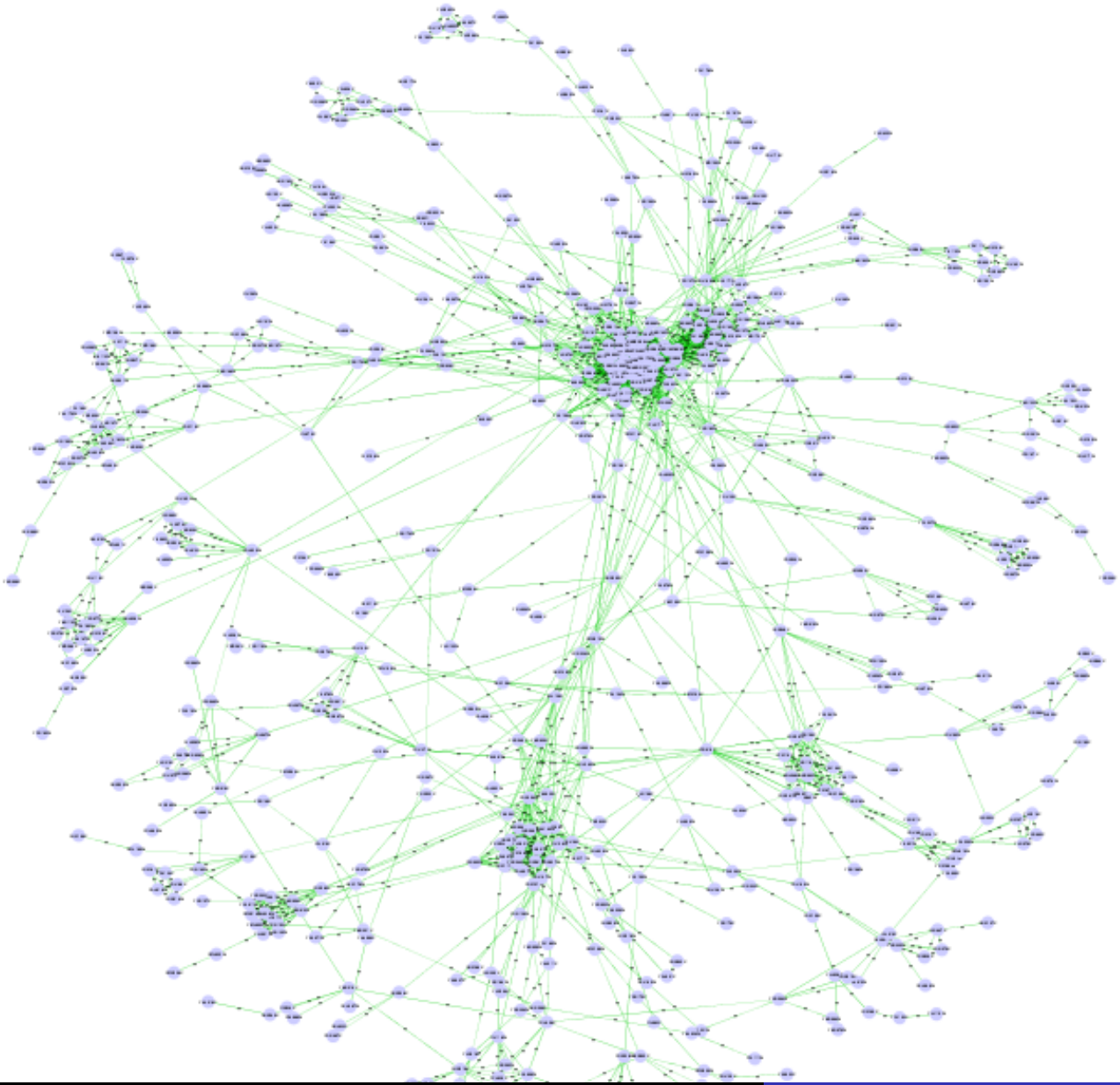


graph and **Rgraphviz**: Graph representation of the integrin mediated cell adhesion pathway with nodes containing pie charts displaying expression levels for the corresponding gene over a set of samples.



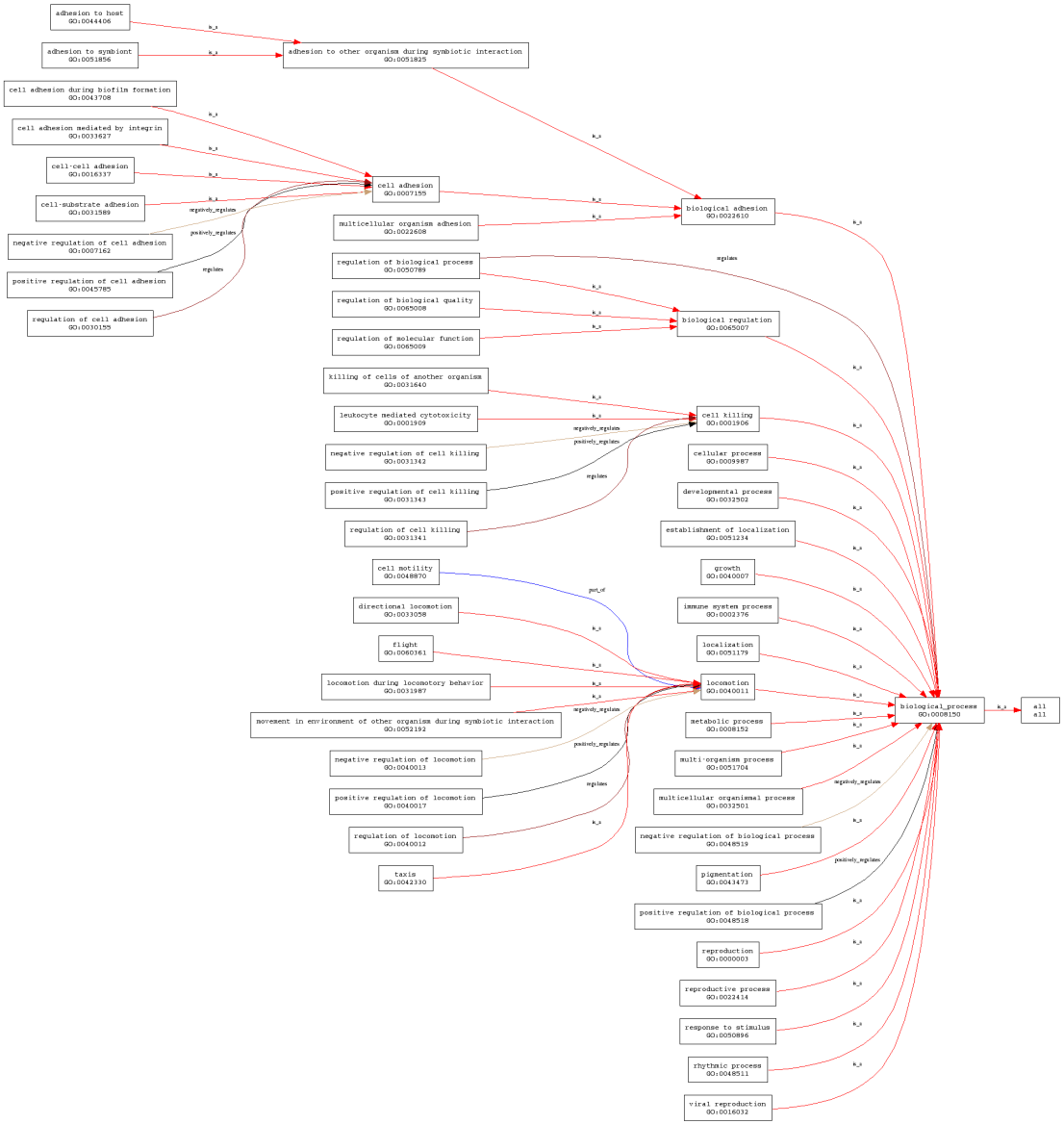
marrayPlots: boxplots of log-ratios log_2 R/G by print-tip group, *maBoxplot()* function.

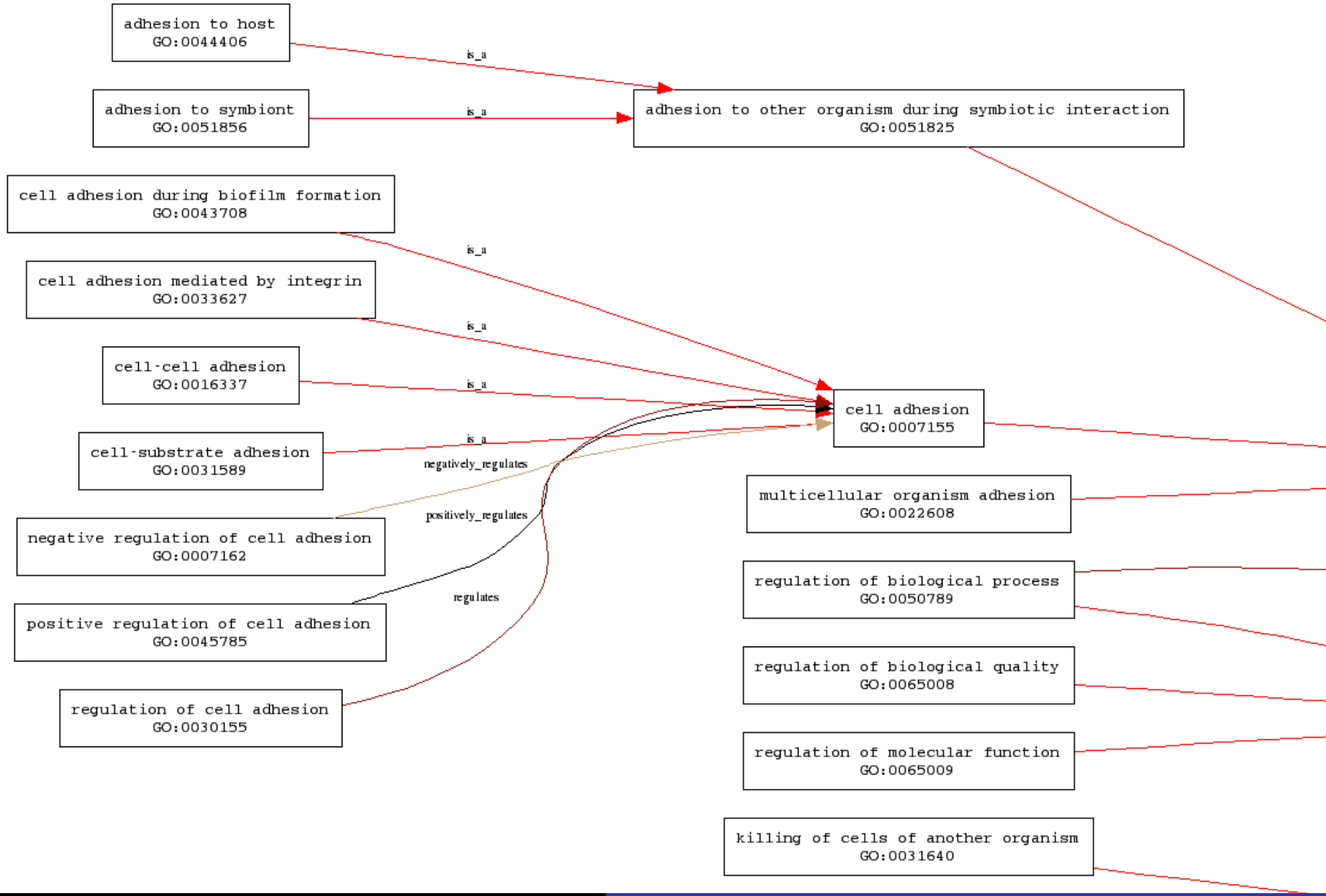
```
source("http://bioconductor.org/biocLite.R")
biocLite()
```



酵母的部分蛋白质相互作用

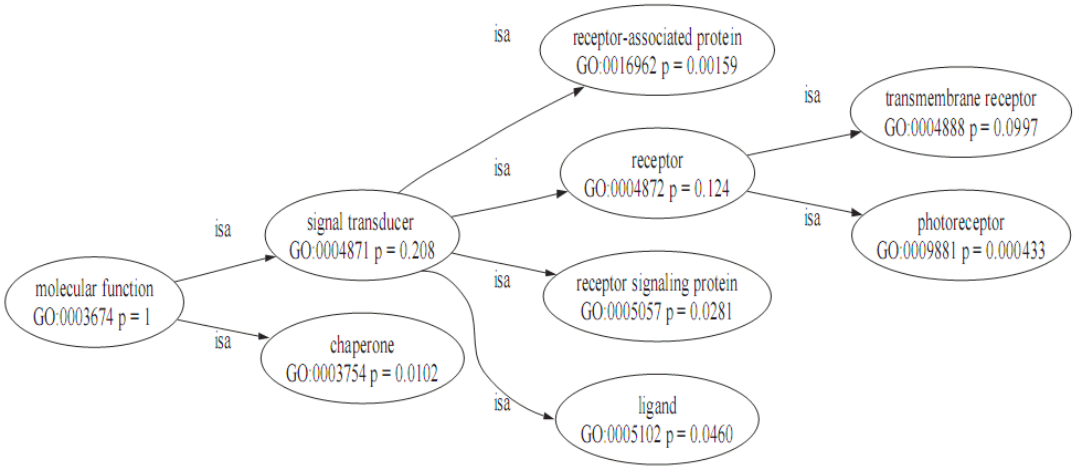
- 点表示蛋白质
- 边表示蛋白质之间的相互作用


adhesion to host
GO:0044406
adhesion to symbiont
GO:0051856
adhesion to other organism during symbiotic interaction
GO:0051825
cell adhesion during biofilm formation
GO:0043708
cell adhesion mediated by integrin
GO:0033627
cell-cell adhesion
GO:0016337
cell-substrate adhesion
GO:0031589
negative regulation of cell adhesion
GO:0007162
positive regulation of cell adhesion
GO:0045785
regulation of cell adhesion
GO:0030155
cell adhesion
GO:0007155
multicellular organism adhesion
GO:0022608
biological adhesion
GO:0022610
regulation of biological process
GO:0050789
regulation of biological quality
GO:0065008
regulation of molecular function
GO:0065009
biological regulation
GO:0065007
killing of cells of another organism
GO:0031640
leukocyte mediated cytotoxicity
GO:0001909
negative regulation of cell killing
GO:0031342
positive regulation of cell killing
GO:0031343
regulation of cell killing
GO:0031341
cell killing
GO:0001906
cellular process
GO:0009987
developmental process
GO:0032502
establishment of localization
GO:0051234
growth
GO:0040007
immune system process
GO:0002376
localization
GO:0051179
cell motility
GO:0048870
directional locomotion
GO:0033058
flight
GO:0060361
locomotion during locomotory behavior
GO:0031987
movement in environment of other organism during symbiotic interaction
GO:0052192
negative regulation of locomotion
GO:0040013
positive regulation of locomotion
GO:0040017
regulation of locomotion
GO:0040012
taxis
GO:0042330
locomotion
GO:0040011
metabolic process
GO:0008152
multi-organism process
GO:0051704
multicellular organismal process
GO:0032501
negative regulation of biological process
GO:0048519
pigmentation
GO:0043473
positive regulation of biological process
GO:0048518
reproduction
GO:0000003
reproductive process
GO:0022414
response to stimulus
GO:0050896
rhythmic process
GO:0048511
viral reproduction
GO:0016032
biological_process
GO:0008150
all
all
is_a
negatively_regulates
positively_regulates
regulates
part_of


adhesion to host
GO:0044406
adhesion to symbiont
GO:0051856
adhesion to other organism during symbiotic interaction
GO:0051825
cell adhesion during biofilm formation
GO:0043708
cell adhesion mediated by integrin
GO:0033627
cell-cell adhesion
GO:0016337
cell-substrate adhesion
GO:0031589
cell adhesion
GO:0007155
negative regulation of cell adhesion
GO:0007162
positive regulation of cell adhesion
GO:0045785
regulation of cell adhesion
GO:0030155
multicellular organism adhesion
GO:0022608
regulation of biological process
GO:0050789
regulation of biological quality
GO:0065008
regulation of molecular function
GO:0065009
killing of cells of another organism
GO:0031640
is_a
negatively_regulates
positively_regulates
regulates

GO Term之间相似性的定义：

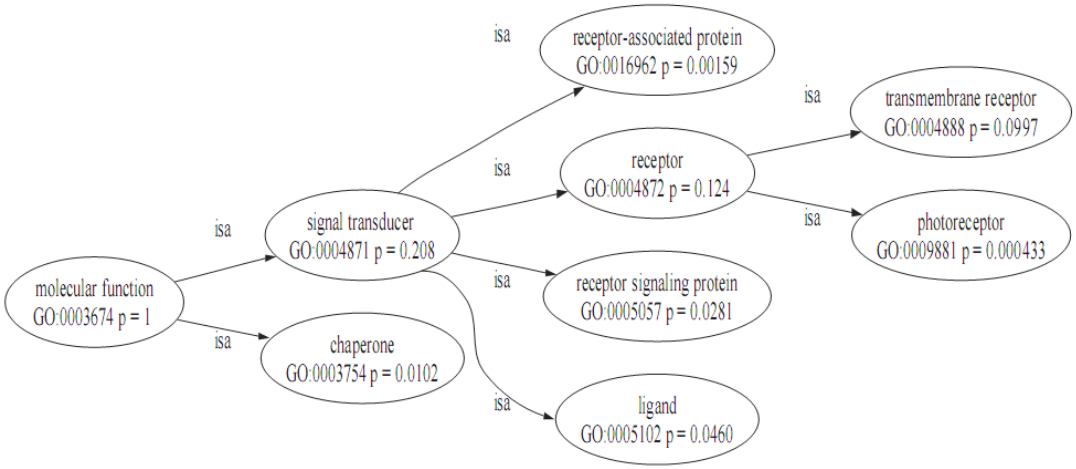
- P. W. Lord



$$Sim_{Lord}(term1, term2) = \min_{term \in (term1, term2)} \{p(term)\}$$

GO Term之间相似性的定义：

- P. W. Lord



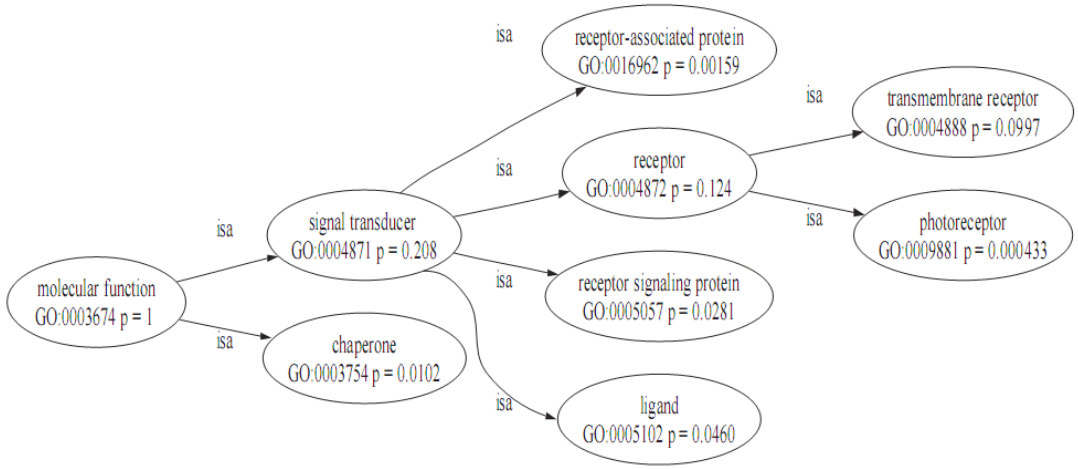
$$Sim_{Lord}(term1, term2) = \min_{term \in (term1, term2)} \{p(term)\}$$

- Resnik

$$Sim_{Resnik}(term1, term2) = -\ln\ Sim_{Lord}(term1, term2)$$

GO Term之间相似性的定义：

- P. W. Lord



$$Sim_{Lord}(term1, term2) = \min_{term \in (term1, term2)} \{p(term)\}$$

- Resnik

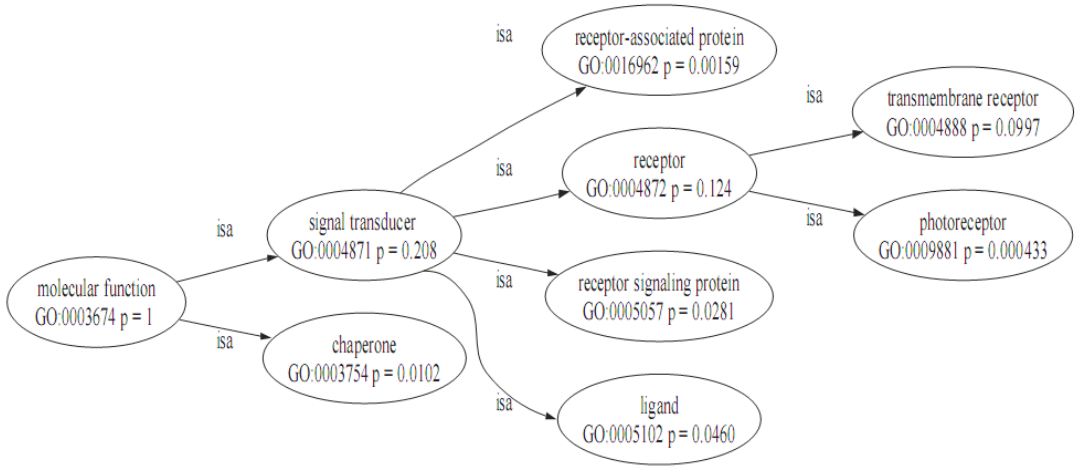
$$Sim_{Resnik}(term1, term2) = -\ln\ Sim_{Lord}(term1, term2)$$

- Jiang

$$Sim_{Jiang}(term1, term2) = 1 - min(1, IC(term1) - 2IC_{ms} + IC(term2))$$

GO Term之间相似性的定义：

- P. W. Lord



$$Sim_{Lord}(term1, term2) = \min_{term \in (term1, term2)} \{p(term)\}$$

- Resnik

$$Sim_{Resnik}(term1, term2) = -\ln\ Sim_{Lord}(term1, term2)$$

- Jiang

$$Sim_{Jiang}(term1, term2) = 1 - min(1, IC(term1) - 2IC_{ms} + IC(term2))$$

蛋白质之间GO语义相似性的定义：

- 平均值

蛋白质之间GO语义相似性的定义：

- 平均值
- 最大值

蛋白质之间GO语义相似性的定义：

- 平均值
- 最大值
- 匹配得最好的GO术语的相似性的平均值

蛋白质之间GO语义相似性的定义：

- 平均值
- 最大值
- 匹配得最好的GO术语的相似性的平均值
- 匹配得最好的GO术语的相似性的最大值

- 数据来源：人类蛋白质参考数据库（Human Protein Reference Database）

- 数据来源：人类蛋白质参考数据库（Human Protein Reference Database）
- 蛋白质数量：25,661

- 数据来源：人类蛋白质参考数据库（Human Protein Reference Database）
- 蛋白质数量：25,661
- 蛋白质相互作用数量：37,107

- GO term的分类：3种

- GO term的分类：3种
- GO Term之间相似性定义：7种

- GO term的分类：3种
- GO Term之间相似性定义：7种
- Gene Product之间的GO相似性定义：4种

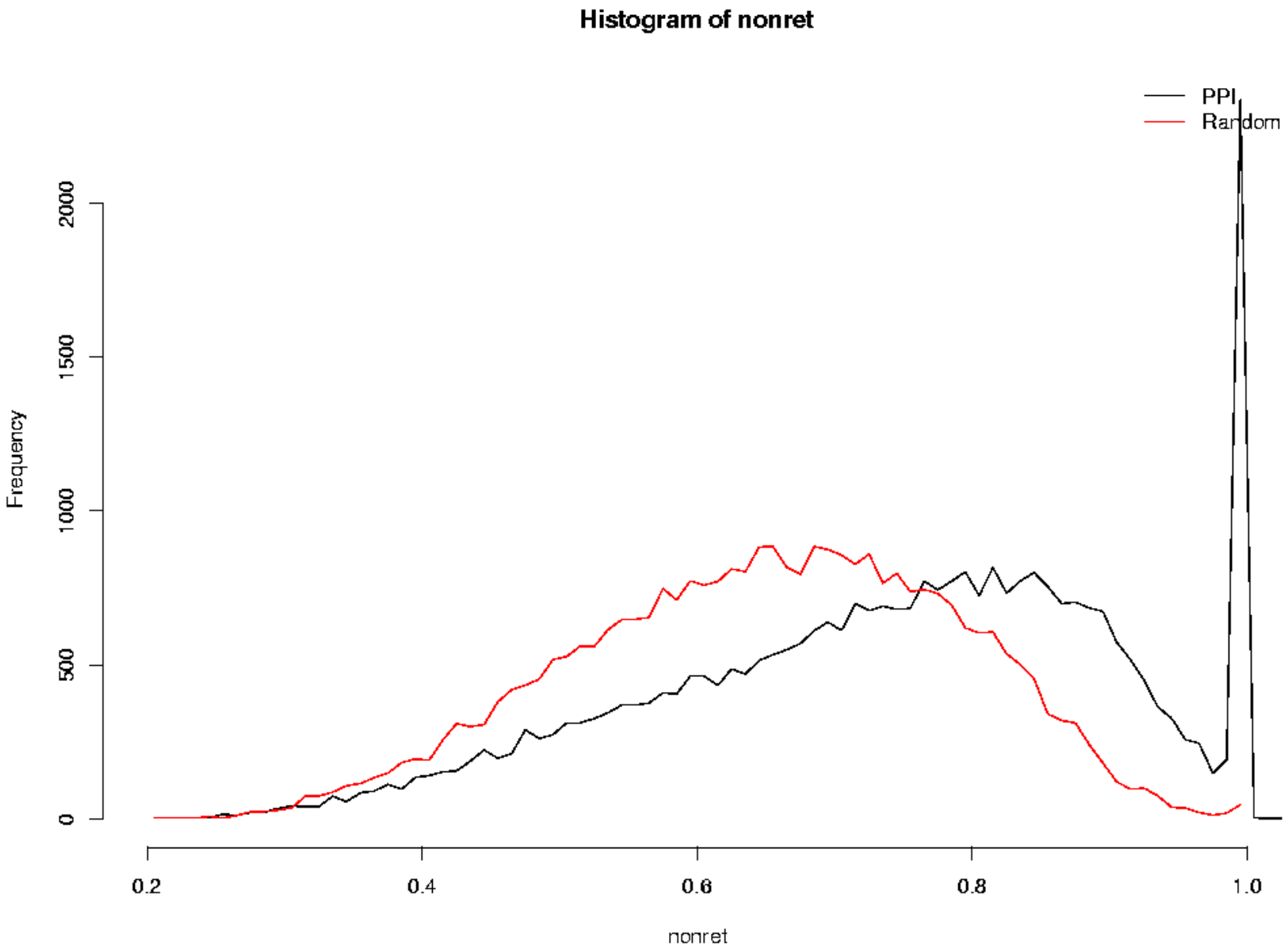
- GO term的分类：3种
- GO Term之间相似性定义：7种
- Gene Product之间的GO相似性定义：4种
- 试验量：$3 \times 7 \times 4 = 84$

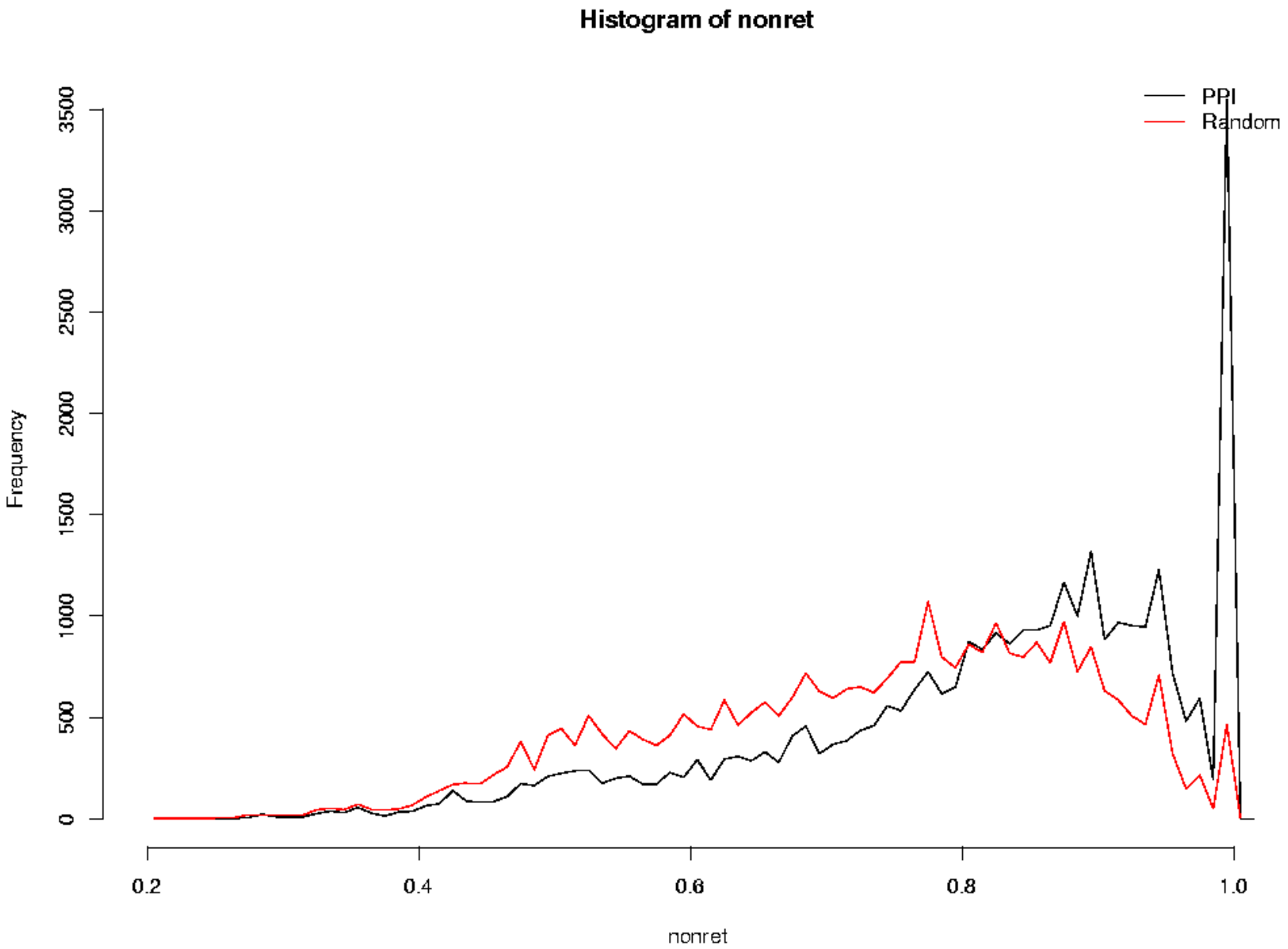
```
filterGenePairs <- function(genePairList) {
    newGeneList <- matrix(ncol=2,nrow=0)

    num <- 1
    for(i in 1:(length(genePairList)/2)){
        if(length(filterGO(genePairList[i,1])) != 0
        && length(filterGO(genePairList[i,2])) != 0 )
            newGeneList <- rbind(newGeneList, genePairList
    }
    return(newGeneList)
}
```

```
getGenePairSim <- function(geneA, geneB, similarity = "funS
similarityTerm = "Lin", verbose = FALSE) {
    if(geneA & FALSE || geneB & FALSE)  return(0)
    if(geneA == geneB)  return (1)

    temp <- getGeneSim(c(geneA, geneB),
                       similarity = similarity,
                       similarityTerm = similarityTerm)
    ret <- temp[2,1]
    return(ret)
}
```


Histogram of nonret
PPI
Random
Frequency
0
500
1000
1500
2000
0.2
0.4
0.6
0.8
1.0
nonret


Histogram of nonret
PPI
Random
Frequency
0
500
1000
1500
2000
2500
3000
3500
0.2
0.4
0.6
0.8
1.0
nonret

	基于Bioconductor的实现	基于Java的实现
代码量	120行	约2500行
运行时间*	20小时	48小时内未完成计算
开发时间**	4小时	约1个月

如何参与？

- 使用、宣传

如何参与？

- 使用、宣传
- 报告Bug

如何参与？

- 使用、宣传
- 报告Bug
- 参软件包的开发，贡献自己的代码

如何参与？

- 使用、宣传
- 报告Bug
- 参软件包的开发，贡献自己的代码
- 贡献自己的软件包

如何参与？

- 使用、宣传
- 报告Bug
- 参软件包的开发，贡献自己的代码
- 贡献自己的软件包
- 进入core team？

所面临的问题:

- R代码的并行？分布式？

所面临的问题:

- R代码的并行？分布式？
- R代码的管理？团队协作？

所面临的问题:

- R代码的并行？分布式？
- R代码的管理？团队协作？
- R的组内培训？

谢谢