基因组学与应用生物学,2013 年,第 32 卷,第 1 期,第 135-141 页
Genomics and Applied Biology, 2013, Vol.32, No.1, 135-141

评述与展望
Review and Progress

基于 WGCNA 算法的基因共表达网络构建理论及其 R 软件实现

宋长新[1*] 雷萍[2] 王婷[1]
1 青海师范大学计算机学院, 青海, 810008; 2 深圳市水质检测中心, 深圳, 518000
* 通讯作者, scx@qhnu.edu.cn

摘 要 WGCNA (weighted gene co-expression network analysis)算法是一种构建基因共表达网络的典型系统生物学算法,该算法基于高通量的基因信使 RNA (mRNA)表达芯片数据,被广泛应用于国际生物医学领域。本文旨在介绍 WGCNA 的基本数理原理,并依托 R 软件包 WGNCA 以实例的方式介绍其应用。WGCNA 算法首先假定基因网络服从无尺度分布,并定义基因共表达相关矩阵、基因网络形成的邻接函数,然后计算不同节点的相异系数,并据此构建分层聚类树(hierarchical clustering tree),该聚类树的不同分支代表不同的基因模块(module),模块内基因共表达程度高,而分数不同模块的基因共表达程度低。最后,探索模块与特定表型或疾病的关联关系,最终达到鉴定疾病治疗的靶点基因、基因网络的目的。

关键词 WGCNA 算法, 基因共表达网络, R 软件

Gene Co-expression Network Analysis Based on WGCNA Algorithm-Theory and Implementation in R Software

Song Changxin[1*] Lei Ping[2] Wang Ting[1]
1 Department of Computer Science, Qinghai Normal University, Qinghai, 810008; 2 Shenzhen Water Quality Centre, Shenzhen, 518000
* Corresponding author, scx@qhnu.edu.cn
DOI: 10.3969/gab.032.000135

Abstract WGCNA (weighted gene co-expression network analysis) is a typical algorithm which is used in gene co-expression network identification. This algorithm is based on high-throughout mRNA gene expression profiles and being widely used in the international biomedical field. In this article, we will introduce the basic theory and it's implementation in R software. Firstly, the scale-free of gene network condition should be satisfied before conducting WGCNA, what's more, it was necessary to define the correlation matrix of gene co-expression and adjacency function. Secondly, the dissimilarity measurements of different nodes were calculated, and then hierarchical clustering tree was built based on these data. Different dendrogram branches represented various modules. There is much higher co-expression strength among genes in the same module than that in different modules. At last, it is critical to connect the modules with interesting phenotypes or disease and identity the target genes for disease treatment.

Keywords WGCNA, Gene co-expression network, R software

基因共表达网络(gene co-expression network)分析致力于寻找协同表达的基因模块(module),并探索基因网络与研究者关注的表型之间的关联关系。它基于高通量的微阵列技术(microarray),应用基因表达芯片得到实验数据,从转录组(mRNA)水平探索基因网与疾病或者研究者关注的性状之间的关联关系,因此应用于复杂疾病的易感基因鉴定、新药开发等生物医学研究领域。

加权基因共表达网络构建(weighted gene co-expression network)算法作为一种高效、准确的生物信息学、生物数据挖掘方法,理论不断完善,应用日渐广泛。

首先,在理论研究上,Zhang 和 Horvath (2005)学者于 2005 年在发表了第一篇理论分析文章,截至到 2012 年 5 月,已有 12 篇 WGCNA 相关的文章先后发表在生物信息、系统生物学领域权威期刊上,如 *BMC Bioinformatics*、*PloS Computational Biology* 和 *Bioin-*

基金项目 本研究由青海省 135 高层次人才培养基金资助

formatics 等。2008 年，基于该研究方法理论，Langfelder 和 Horvath (2008)编写并发布了 WGCNA R 软件包，截至到 2012 年 4 月引用次数达到 108 次之多。

而在实际应用层面，由于基因共表达网络分析作为一种挖掘和呈现基因在不同样本中表达形式的有效方法，可以鉴定高度共表达的基因模块，模块的特征值或模块中包含的关键基因(hup genes: 即与基因网络中的其它基因联系最紧密，在基因网中起关键作用的基因)可用于提炼模块信息，以此深入探讨基因模块或模块中的关键基因和研究人员关注的样本特征间的关联关系。

WGCNA 算法已被用于鉴定复杂疾病的候选标记或药物靶点，并用于多项人类复杂疾病的研究。如鉴定家族性混合型高脂血症(Plaisier et al., 2009)、胶质母细胞瘤(Horvath et al., 2006)、自闭症(Voineagu et al., 2011)、阿尔兹海默病(Miller et al., 2010)、骨质疏松症(Farber, 2010)的关联基因、生物学通路和肿瘤治疗靶点。

Horvath 等(2006)学者将 WGCNA 方法用于胶质母细胞瘤的研究，充分利用两组共 120 个患者的胶质母细胞的基因表达数据，挖掘得到一个包含的基因与已知的癌症相关模块“metasignature”中基因高度重叠的基因共表达模块，且异常纺锤型小脑畸形症相关基因(abnormal spindle-like microcephaly associated, ASPM)为该模块中的一个关键基因，ASPM 基因已被证实为一个胶质母细胞瘤治疗的靶点基因。

Farber (2010)应用 12 名低骨密度和 14 名高骨密度绝经前妇女的单核细胞 mRNA 基因表达数据，使用 WGCNA 方法构建表达共基因模块，并发现模块 9 和骨密度存在显著关联关系，且该结论得到了弗莱明翰(Kiel et al., 2007)基于家系和群体数据的基于单核苷酸多态性(SNP)的全基因组关联研究(GWAS)结果和 dCOD 基因遗传学研究(Styrkarsdottir et al., 2008)的 GWAS 结果的支持。

综上所述，深入了解 WGCNA 算法的基本原理，并掌握该方法，将其运用到实际的科学研究中具有极其重要的意义。本文首次在中文期刊上详细介绍 WGCNA 算法，旨在在国内推广该系统生物学研究方法。

1 原理介绍

WGCNA 算法是构建基因共表达网络的常用算法。首先基因共表达网络的概念为：每个节点代表一个基因，在不同样本中存在表达共性的基因处于同一个基因网络，而基因间的共表达关系一般由它们之间的表达相关系数衡量。一般而言，同一基因共表达网络中的基因表达形式相似、而不同基因共表达网络中的基因表达形式差别较大。构建了基因模块后，再将基因模块与一些表型信息，例如是否患病、身高和体重等联系起来，以探索基因模块形成的根本原因，鉴定出与表型相关的基因模块。以下具体介绍 WGCAN 算法(陈超, 2011)。

1.1 网络构建前提

在 WGCNA 算法中，定义的基因共表达矩阵中的元素为基因的相关系数的加权值，权重的选择标准为使得每个基因网中包含的基因之间的连接服从无尺度网络分布(scale-free networks) (Barabási, 2009)，即连接数为 i 的概率 p(i)与 i 的 n 次方成反比，即 $p(i)\sim i^{-r}$。在实际应用中研究者通过选择加权系数来逼近无尺度网络分布，使之满足如下条件：连接节点个数的对数(log(i))与此节点出现概率的对数值(log(p(i)))为负相关关系(注意：相关系数至少应达到 0.8)，与此同时，不同模块中的基因的平均连接度应比较高。

1.2 网络构建步骤

1.2.1 基因共表达相关矩阵的定义

基因共表达相关矩阵中的元素为基因之间的两两相关系数，即每对基因 m 和基因 n 的相关系数为 $S_{mn}=|cor(m,n)|$，据此构成了基因共表达相关矩阵 $S=[S_{mn}]$。

1.2.2 定义邻接函数

最直接的邻接函数通过指定基因之间的相关系数的阈值(如 $R=0.85$)将基因对划分为相关的和不相关的，这种分法虽然简单易行，但将损失大量信息，如将阈值定为 0.85 时，即便是相关系数为 0.84 的基因对也将被划分到“不相关”的组中。鉴于此，WGCNA 算法中应用了幂指数邻接函数。即对于任何基因对，用邻接系数 a_{mn} 作为衡量它们之间相关关系的指标：$a_{mn}=power(S_{mn},\beta)=|S_{mn}|^{\beta}$，即对相关系数进行 次方的幂指数加权。

1.2.3 确定邻接函数的参数

根据无尺度网络原则确定加权系数 β，即：连接节点个数取对数(log(k))和节点出现的概率的对数值(log(p(k)))之间的相关系数至少达到 0.8。

1.2.4 节点间的相异度衡量

在确定了邻接函数参数 β 后，接下来将相关矩阵 $S=[S_{mn}]$ 转换成邻接矩阵 $A=[a_{mn}]$。将某个基因和分析中其他所有基因之间的关系纳入考虑，邻接矩阵被转换成拓扑矩阵 $\Omega=[\omega_{mn}]$ (Ravasz et al., 2002)，矩阵中的元素如下所示：

$$\omega_{mn}=\frac{l_{mn}+a_{mn}}{\min\{k_m,k_n\}+1-a_{mn}}$$

公式中 $l_{mn}=\sum_\mu a_{m\mu}a_{\mu n}$ 代表和基因 m、n 都存在连接的节点的邻接系数乘积和，$k_m=\sum_\mu a_{m\mu}$ 为基因 m 单独连接的节点的邻接系数加和，类似地，$k_n=\sum_\mu a_{n\mu}$ 代表基因 n 单独连接的节点的邻接系数加和。在基因 m 和基因 n 之间无连接，且无任何其它的基因将这两个基因连接的情况下 $\omega_{mn}=0$。与此同时，节点的相异程度用 $d^{\omega}_{mn}=1-\omega_{mn}$ 来衡量，d^{ω}_{mn} 是网络构建的基础。

1.2.5 聚类分析鉴定基因模块

以基因之间的相异系数 d^{ω}_{ij} 为基本元素构建分层聚类树(hierarchical clustering tree)，聚类数的不同分支代表不同的基因模块。构建聚类树有静态剪切树和动态剪切树两种算法(Langfelder et al., 2008)。简而言之，静态剪切树是通过定义一个固定的高度将群集分支，该方法识别群集的准确度不高。而动态剪切算法基于树状图的分支形状，它可挖掘得到静态剪切无法检测出的基因模块，更重要的是，用动态剪切算法鉴定出的基因网络以往的生物实验结果比较一致(Dong and Horvath, 2007)。两种动态剪切树算法简介如下：(1)动态树法，为一种“自上而下”的算法，由不断的分解和组合这一反复迭代过程，在迭代结果稳定后，鉴定出基因模块，该算法由根据静态剪切法得到的几个较大的基因模块开始，对每个群集加入波动特征模式进行精细的分析和分割；(2)动态混合切割法，是一种“自下而上”的算法，主要由识别和测试步骤组成，首先识别满足如下条件的基因模块：(a)模块中基因个数满足设定的最低数目；(b)从固定的基因模块中移除距离过远的分支；(c)基因模块间的区别明显；(d)基因模块中的关键基因(hup gene)应紧紧连接；测试阶段为测试未分配的基因，将其分配进足够接近的初步模块，并最终形成基因网络/模块结构。

基因模块与已知特征相联系的具体办法有：(1)计算得到基因网络/模块的特征值，再计算模块的特征向量与关注表型的相关系数；(2)对于分组表型数据(如疾病状态等)，可以首先定义应用 T 检验计算每个基因在不同组(如疾病组和正常组)间的基因 mRNA 差异表达的显著性检验 P 值，并将显著性 P 值的以 10 为底的对数值定义为基因显著性(GS: gene significance)，再将每一个模块显著性(module significance, MS)定义为模块中所包含基因的 GS 的平均值。然后比较 MS 的值，一般而言，某模块的 MS 值显著高于其它模块说明这一个模块可能与疾病存在关联关系；(3)利用基因网络中的包含和其它基因连接度较高的关键基因的信息来推测该基因网络/模块的成因。

2 R 软件实例

本小节将详细介绍在 R 软件应用中(R 2.3.12, http://www.r-project.org/)，应用 WGCNA 软件包(Langfelder and Horvath, 2008)的实例。具体为生成模拟基因表达谱数据以及表型数据，构建基因模块，并探索其与表型之间的关联关系。有关 WGCNA 的最新研究进展参见其主页 http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/。

2.1 模拟数据

首先在 WGCNA 算法主页下载 WGCNA Package 并将其安装在 PC 上，设置好程序运行的工作空间，并加载程序包，然后模拟基因 mRNA 表达谱以及样本表型数据，得到 30 个样本的 3 600 个基因的表达数据和样本的患病状态，用变量 y 表示(y=0 代表正常组，y=1 代表患病组)。具体的 R 程序如下所示(符号“#”后为注释)：

```
### 模拟数据
setwd("E:/WGCNA/simulate")        # 设定工作空间，请根据实际情况设置
library("WGCNA")                  # 加载程序包
options(stringsAsFactors=FALSE)   # 预设置
# 步骤一 预设参数
no.obs<-30                        # 设置样本个数
ESturquoise<-0; ESbrown<--.6; ESgreen<-.6; ESyellow<-0   # 设置基因模块的显著性水平
ESvector<-c (ESturquoise, ESbrown, ESgreen, ESyellow)
nGenes1<-3000                     # 设置基因个数为 3 000
simulateProportions1 <-c (0.2, 0.15, 0.08, 0.06, 0.04)   # 模块中基因个数比
set.seed (1)
# 步骤二 模拟基因网
MEgreen<-rnorm (no.obs)
scaledy<-MEgreen*ESgreen+sqrt (1-ESgreen^2)*rnorm (no.obs)
y<-ifelse(scaledy>median (scaledy), 1, 0)
# 模拟的表型变量
MEturquoise<-ESturquoise*scaledy+sqrt (1-ESturquoise^2)*rnorm (no.obs)
MEblue<-.6*MEturquoise+sqrt (1-.6^2)*rnorm (no.obs)
MEbrown<-ESbrown*scaledy+sqrt (1-ESbrown^2)*rnorm(no.obs)
MEyellow<-ESyellow*scaledy+sqrt (1-ESyellow^
```

```
2)*rnorm(no.obs)
    MEN1<-data.frame(y, MEturquoise, MEblue, MEbrown, MEgreen, MEyellow)
    dat1<-simulateDatExpr5Modules(MEturquoise=MEN1$MEturquoise,
    MEblue=MEN1$MEblue, MEbrown=MEN1$MEbrown, MEyellow=MEN1$MEyellow,
    MEgreen=MEN1$MEgreen, nGenes=nGenes1,
    simulateProportions=simulateProportions1)
    datExpr<-dat1$datExpr;
    truemodule<-dat1$truemodule;
    datME<-dat1$datME;
    attach(MEN1)
   datExpr<-data.frame(datExpr)
    #模拟的表达谱数据
    ArrayName<-paste("Sample",1:dim(datExpr)[[1]], sep="")
    GeneName<-paste("Gene",1:dim(datExpr)[[2]], sep="")
    Dimnames(datExpr)[[1]]=ArrayName
    Dimnames(datExpr)[[2]]=GeneName
```

为考察数据中是否包含有离群的样本，我们提交如下程序绘制样本聚类图，并且将各样本的患病状态标注在图中：

```
    plotClusterTreeSamples(datExpr=datExpr, y=y)
    #绘制样本聚类图。
```

观察样本聚类图(图 1)发现模拟数据中并未包含有离群样本。图中下半部分中样本对应为红色的表示其属于患病组，而黑色部分代表正常组。

2.2 网络构建

2.2.1 选择构建网络参数

为了尽量满足无尺度网络分布前提条件，编写如下 R 程序探索邻接矩阵权重参数 β 的取值：

```
    #设置网络构建参数选择范围，计算无尺度分布拓扑矩阵
    powers=c(c(1:10), seq(from=12, to=20, by=2))
sft=pickSoftThreshold(datExpr, powerVector=powers, verbose=5)
    #图形绘制
    parb(mfrow=c(1,2));
   cex1=0.9;
    plotb(sft$fitIndices[,1], -signb(sft$fitIndices[,3])*sft$fitIndices[,2],
    xlab="Soft Threshold (power)", ylab="Scale Free Topology Model Fit, signed R^2", type="n",
    main=paste("Scale independence"), ylim=c(0,1));
    text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
    labels=powers,cex=cex1, col="black");
    abline(h=0.90, col="black")
    plot(sft$fitIndices[,1], sft$fitIndices[,5], type="n", main=paste("Mean connectivity"),
    ylab="MeanConnectivity", xlab="SoftThreshold (power)")
    text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1, col="black")
```

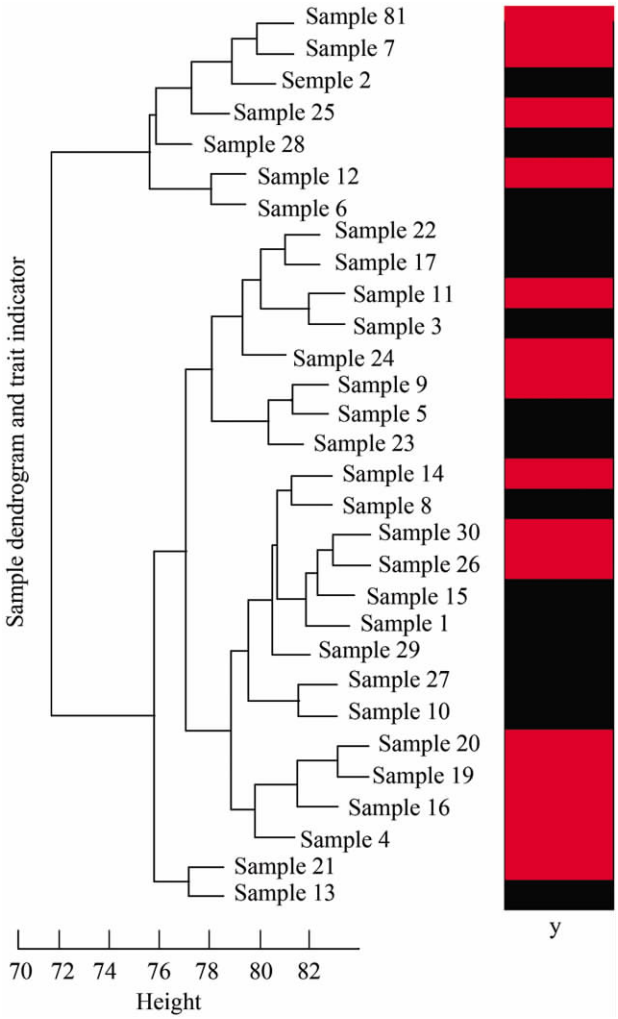

图 1 系统聚类树的结果与对应的样本信息

Figure 1 The phylogenetic tree and the information of the sample

使参数 β 取值从 1~18，计算相应的模型选择统计量绘制图形(图 2)，这两个分图的横轴均代表权重参数 β，左图的纵轴代表对应的网络中 log(k)与 log(p(k))相关系数的平方。相关系数的平方取值越高(最小应达到 0.8)，说明该网络越逼近无网络尺度的分布。B 图的纵轴代表对应的基因模块中所有基因邻接函数的均值。本研究发现在模拟的数据中，选择 log(k)与 log(p(k))的相关系数的平方取值均比较高，因此我们选择其值

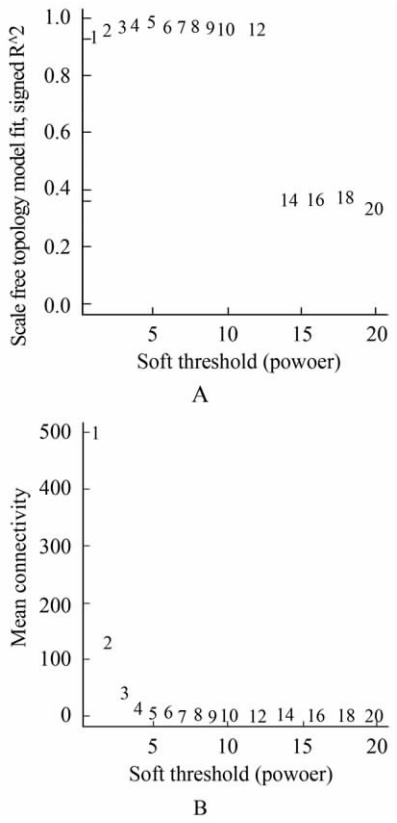

图 2 β 参数选择图

注: A: **刻度独立值**; B: **平均连通性**

Figure 2 Figures for select β

Note: A: Scale independence; B: Mean connectivity

首次接近 0.90 时 β 的取值，即 β=2，据此构建基因网络。

2.2.2 构建网络

首先计算基因间的相异性系数，并得到系统聚类树，然后按照混合动态剪切树的标准，并设置每个基因网络 / 模块最少的基因数目为 30。在使用动态剪切法在确定基因模块后，我们依次计算每个模块的特征向量值(eigengenes)**，然后对模块进行聚类分析，将距离较近的模块合并成新的模块**(**可设置** height=0.2)**。本研究最终生成的模块聚类图如图 3 所示。图中不同颜色代表不同的基因模块，而灰色部分代表无法合并到任何其它模块中的基因。本步骤的 R 程序如下所示：**

```
softPower<-2;                    # 设定无尺度参数为 2
adjacency = adjacency (datExpr, power=softPower);
TOM=TOMsimilarity (adjacency);
dissTOM=1-TOM
# 聚类分析
geneTree=flashClust (as.dist (dissTOM), method="average");
minModuleSize=30;
# 设置基因网中至少包含 30 个基因
dynamicMods=cutreeDynamic (dendro=geneTree, distM=dissTOM,
    deepSplit=2, pamRespectsDendro=FALSE,
    minClusterSize=minModuleSize);
    dynamicColors=labels2colors (dynamicMods)
    table (dynamicColors)
    # 计算基因网的特征值
    MEList=moduleEigengenes (datExpr, colors=dynamicColors)
    MEs=MEList$eigengenes
    MEDiss=1-cor (MEs);
    METree=flashClust(as.dist (MEDiss), method="average");
    # 将特征值距离小于 0.2 的基因网合并
    MEDissThres=0.2
    merge=mergeCloseModules (datExpr,dynamicColors,cutHeight=MEDissThres,verbose=3)
    mergedColors=merge$colors;
    mergedMEs=merge$newMEs;
    ## 重新命名合并后的基因网
    moduleColors=mergedColors
    colorOrder=c ("grey", standardColors (50));
    moduleLabels=match(moduleColors,colorOrder)-1;
    MEs=mergedMEs;
    ## 绘制最终基因网络构建图
    plotDendroAndColors (geneTree, mergedColors, "Merged dynamic",
    dendroLabels=FALSE, hang=0.03,
    addGuide=TRUE, guideHang=0.05)
```

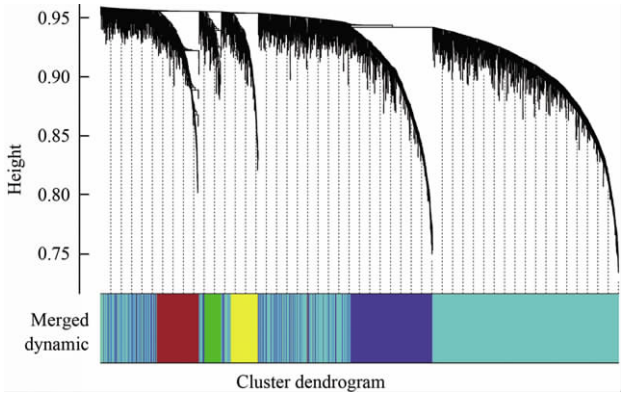

图 3 基因的系统聚类数和动态剪切法生成的基因网络 / 模块

Figure 3 Cluster dendrogram based on dynamic tree cut

2.3 网络与表型关联分析

2.3.1 相关系数法

计算每个基因模块的特征值与患病状态(变量 y**)的相关系数，编写程序如下所示：**

```
Signif (cor(y, datME, use="p"), 2)
pvalues=corPvalueStudent (cor (y, datME, use="p"), nSamples=length(y))
```

得到结果见表 1。

表 1 基因模块特征值与表型相关系数

Table 1 Correlation coefficient between model eigengenes and trait

模块 Model	MEturquoise	MEblue	MEbrown	MEyellow	MEgreen
系数 Coefficient	0.44	0.19	-0.57	-0.12	0.41
p 值 p value	0.01	0.31	<0.01	0.518	0.03

以上结果表明，模块 brown 和本的患病状态存在较强的负相关关系(R=-0.57, p<0.01)。联系本研究模拟数据的实际背景，即模块 brown 可能与样本的患病状态存在关联关系，且该模块中的基因表达对疾病可能产生抑制作用。

2.3.2 基因网显著性

编写如下程序计算每个基因模块的 GS 值，并且绘制相应图形：

```
# 计算基因的显著性
GS1=as.numeric (cor (y, datExpr, use="p"))
GeneSignificance=abs (GS1)
# 计算基因模块的显著性
ModuleSignificance=tapply (GeneSignificance, moduleColors, mean, na.rm=T)
# 图形绘制
plot ModuleSignificance (GeneSignificance, moduleColors, ylim=c (0,0.5), main="Module Significance")
```

观察图 4 可知，模块 brown 的 MS 值为所有的基因模块中最高的(>0.3)，同样说明该模块可能与疾病存在关联关系，与相关分析的结果一致。

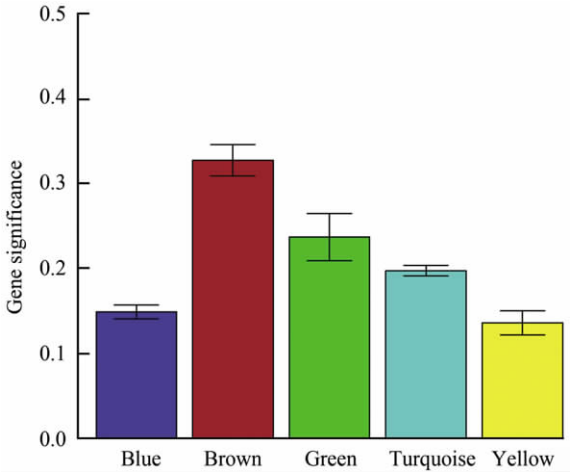

图 4 不同的基因模块 GS 取值均值误差图

Figure 4 Error bars of gene significance for different model

3 结论与展望

WGCNA 算法通过构建基因模块来从基因表达芯片数据中挖掘出有效信息，并力求从生物学角度解释基因模块的意义。由于组织、细胞中的基因转录本表达极其复杂，虽然先进的生物学技术让我们能够借助于高通量基因表达芯片，得到基因的 mRNA 表达数据，但从众多的的基因 mRNA 表达水平中提取信息仍然是生物信息、数据挖掘领域的难题。在 WGCNA 算法开创前，基于相关系数的基因共表达网络分析中最大的问题在于将基因间的关系生硬地划分为“相关”和“不相关”(Carter et al., 2004)，例如设定相关系数的阈值来判断基因之间是否存在关联，从而导致信息损耗。而 WGCNA 算法相对以往算法最大的优势为在分析前假设基因之间的相关关系是符合无尺度网络分布的。已有生物学研究表明某些关键基因在一些生理过程中所起的作用为不可替代的，而某些基因在生物学过程中所执行的功能被限制在了某些特定的信号通路中间，以上发现和无尺度网络分布的概念是契合的。

WGCNA 算法通过选择合适的加权系数对基因之间的相关系数加权，使之满足基因网近似服从无尺度网络分布。由此可见，基因网络 / 模块的获得，是通过经过矩阵变换后构建的系统聚类树实现，这样保证了基因网络 / 模块的构建为软阈值，即相关系数被当成连续变量处理。再通过动态剪切树的算法对基因进行聚类，具有高度表达相似性的基因将会分配到同一个基因网络 / 模块，此后将作为一个整体，探索其与疾病或者相关性状之间的关联关系。与此同时，基因共表达网络分析相对于以 SNP、基因为单位的 GWAS 分析拥有更好的统计功效，因为避免了大量的多重校正导致的假阴性结果。例如，在本研究中，若以探针为单位进行分析与检验，需要进行 3 600 次检验，而应用基因共表达网络分析将数据维度降至 5 个模块，从而将基因与疾病的关联分析转换为基因模块与疾病的关联分析。更重要的是，网络分析提供了基因之间的信息，具体而言，我们鉴定了一个基因集合(模块)，它们在人群中拥有类似的表达形式、有协同表达的特征。而在得到了基因模块后，借助于 DAVID 等生物信息学工具对模块中的基因进行功能注释，探索其具体的生物学意义。而此类分析已在不少研究中得到实践(Aggarwal et al., 2006; Nayak et al., 2009; Dewey et al., 2010; Childs et al., 2011; Ficklin and Feltus, 2011; Zheng et al., 2011)。

WGCNA 虽然得到了广泛的应用，但我们期待它在如下方面得到改进：首先，增加计算运行效率，如开发基于图像处理器(GPU)的软件包等，再者现有算法综合考虑计算时间和内存，建议分析的最大基因个数为 3 600，而随着技术的发展和分析实际需

求,期望改进算法以适用于更多的分析基因。而对于研究者而言,借助于 WGCNA 算法构建了基因模块,为了更深层的挖掘基因模块的信息,可以根据研究目的酌情对感兴趣的基因模块中的关键基因进行基因的功能研究。

作者贡献

宋长新负责写作和修改论文;霭萍和王婷参与文献研究和主题讨论。

致谢

本研究由青海省 135 高层次人才培养基金资助。

参考文献

Aggarwal A., Guo D.L., Hoshida Y., Yuen S.T., Chu K.M., So S., Boussioutas A., Chen X., Bowtell D., Aburatani H., Leung S.Y., and Tan P., 2006, Topological and functional discovery in a gene coexpression meta-network of gastric cancer, Cancer Res., 66(1): 232-241

Barabási A.L., 2009, Scale-free networks: A decade and beyond, Science, 325(5939): 412-413

Carter S.L., Brechbühler C.M., Griffin M., and Bond A.T., 2004, Gene co-expression network topology provides a framework for molecular characterization of cellular state, Bioinformatics, 20(14): 2242-2250

Chen C., 2011, Evaluation of six batch adjustment methods in expression microarray data and application of gene co-expression network in schizophrenia, Dissertation for Ph.D., Fudan University, Supervisor: Jin L., pp.81-86 (陈超, 2011, 基因表达谱芯片校正批次效应算法的比较及网络分析在精神分裂症研究中的应用, 博士学位论文, 复旦大学, 导师: 金立, pp.81-86)

Childs K.L., Davidson R.M., and Buell C.R., 2011, Gene coexpression network analysis as a source of functional annotation for rice genes, PLoS One, 6(7): e22196

Dewey F.E., Perez M.V., Wheeler M.T., Watt C., Spin J., Langfelder P., Horvath S., Hannenhalli S., Cappola T.P., and Ashley E.A., 2010, Gene coexpression network topology of cardiac development, hypertrophy, and failure, Circ. Cardiovasc. Genet., 4(1): 26-35

Dong J., and Horvath S., 2007, Understanding network concepts in modules, BMC Syst. Biol., 1: 24

Farber C.R., 2010, Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data, J. Bone Miner. Res., 25(11): 2359-2367

Ficklin S.P., and Feltus F.A., 2011, Gene coexpression network alignment and conservation of gene modules between two grass species: Maize and rice, Plant Physiol., 156(3): 1244-1256

Horvath S., Zhang B., Carlson M., Lu K.V., Zhu S., Felciano R. M., Laurance M.F., Zhao W., Qi S., Chen Z., Lee Y., Scheck A.C., Liau L.M., Wu H., Geschwind D.H., Febbo P. G., Kornblum H.I., Cloughesy T.F., Nelson S.F., and Mischel P.S., 2006, Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target, PNAS, 103(46): 17402-17407

Kiel D.P., Demissie S., Dupuis J., Lunetta K.L., Murabito J.M., and Karasik D., 2007, Genome-wide association with bone mass and geometry in the framingham heart study, BMC Med. Genet., 8(Suppl 1): S14

Langfelder P., and Horvath S., 2008, WGCNA: An R package for weighted correlation network analysis, BMC Bioinformatics, 9: 559

Langfelder P., Zhang B., and Horvath S., 2008, Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R, Bioinformatics, 24(5): 719-720

Miller J.A., Horvath S., and Geschwind D.H., 2010, Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways, PNAS, 107(28): 12698-12703

Nayak R.R., Kearns M., Spielman R.S., and Cheung V.G., 2009, Coexpression network based on natural variation in human gene expression reveals gene interactions and functions, Genome Res., 19(11): 1953-1962

Plaisier C.L., Horvath S., Huertas-Vazquez A., Cruz-Bautista I., Herrera M.F., Tusie-Luna T., Aguilar-Salinas C., and Pajukanta P., 2009, A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia, PLoS Genet., 5(9): e1000642

Ravasz E., Somera A.L., Mongru D.A., Oltvai Z.N., and Barabási A.L., 2002, Hierarchical organization of modularity in metabolic networks, Science, 297(5586): 1551-1555

Styrkarsdottir U., Halldorsson B.V., Gretarsdottir S., Gudbjartsson D.F., Walters G., Ingvarsson T., Jonsdottir T., Saemundsdottir J., Center J.R., Nguyen T.V., Bagger Y., Gulcher J.R., Eisman J.A., Christiansen C., Sigurdsson G., Kong A., Thorsteinsdottir U., and Stefansson K., 2008, Multiple genetic loci for bone mineral density and fractures, N. Engl. J. Med., 358(22): 2355-2365

Voineagu I., Wang X., Johnston P., Lowe J.K., Tian Y., Horvath S., Mill J., Cantor R.M., Blencowe B.J., and Geschwind D.H., 2011, Transcriptomic analysis of autistic brain reveals convergent molecular pathology, Nature, 474(7351): 380-384

Zhang B., and Horvath S., 2005, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol., 4: Article17

Zheng X., Liu T., Yang Z., and Wang J., 2011, Large cliques in *Arabidopsis* gene coexpression network and motif discovery, J. Plant Physiol., 168(6): 611-618