第 34 卷 第 2 期
2015 年 3 月

华 中 农 业 大 学 学 报
Journal of Huazhong Agricultural University

Vol. 34 No. 2
Mar. 2015,101～105

基于 Copula 函数的权重基因共表达网络分析法

汪伟平[1,2] 白 婷[3] 贺 帅[1,2] 李 倩[4] 胡动刚[5]

1. 北京师范大学地表过程与资源生态国家重点实验室,北京 100875;
2. 民政部/教育部减灾与应急管理研究院,北京 100875;
3. 武汉大学测绘遥感信息工程国家重点实验室,武汉 430079;
4. 华中农业大学生命科学技术学院,武汉 430070; 5. 华中农业大学理学院,武汉 430070

摘要 利用 Frank-Copula 函数构建相关系数,并应用于权重基因共表达网络分析模型的改进。同时,利用该改进模型分析小鼠基因数据,发现在相关度最大的模块中有 67 个基因与小鼠肥胖导致的体质量问题相关。其中,模型的筛选结果有效率为 62.6%,说明其在功能基因筛选应用前景中的科学性、有效性和可行性。

关键词 权重基因共表达网格分析; 相关系数; Copula 函数; 小鼠体质量; 基因筛选

中图分类号 R 857.3 **文献标识码** A **文章编号** 1000-2421(2015)02-0101-05

权重基因共表达网络分析(weighted gene co-expression network analysis,WGCNA)是由 Zhang 等[1]在 2005 年第一次系统地提出的基因筛选方法。与传统网络分析方法不同的是,权重基因共表达网络分析不需要大量信息,仅依靠增加样本数量即可提高精度。值得注意的是,权重基因共表达网络分析应用了生物新陈代谢网络中较为先进的无标度特性理论(scale-free)构建模型,其先进性在阿兹海默症[2]、慢性癫痫[3]和慢性疲劳综合征[4]等疾病的致病基因筛选应用中均已经得到证明。

然而,权重基因共表达网络分析在衡量变量之间的相关关系方面存在不足,其在测度 2 个基因表达值之间的相关关系过程中,Person 只能确定线性关系以及非线性变换下单调性不变的关系,不仅要求数据近似服从正态分布,而且必须通过基于 Fisher 变换的精确检验和近似分布检验后方可采用;此外 Spearman 相关系数在处理尾部有明显的外形轮廓偏离的样本数据时不够敏感[5]。因此,本文通过使用 Copula 函数测度 2 个基因表达值之间的相关关系来改进权重基因共表达网络分析模型,使其能够适应不同条件下的基因共表达数据,进而构建加权关联网络并获得更符合生物意义的结果。

1 Copula 函数原理

1959 年,Sklar[6]首次提出 Copula 函数并应用于数理统计学中,表示将一系列一维的边缘分布函数连接起来构造成多维联合分布函数的函数。Copula 函数将边缘分布和变量间的相关结构分开考虑,使其比较适合于研究具有不同分布特征随机变量之间的相关性分析,克服了传统相关性研究受到边缘分布的影响,即:相关系数不能取[-1,1]区间内的任意值[7],有助于更好地衡量变量之间的相关结构和关系。同时,根据 Copula 函数导出的相关性测度指标还可以衡量变量之间的非线性相关关系,且对于严格单调递增的变换保持不变,并能够获得随机变量之间的相关程度和相关模式[8]。因此,使用 Copula 函数提供了比其他相关性测度指标更多的信息,可以更为全面地体现随机变量之间的相关关系,这是一种更为合理的相关性测度指标方法。

1.1 基于 Copula 函数的相关性测度指标

应用 Copula 函数,可以将相关程度和相关模式的研究有机地结合起来,较好地度量变量之间的相关关系。其中,本文应用 Frank-Copula 函数,因为其更具对称性且密度分布呈 U 型,适于描述具有对称相关结构变量之间的相关关系[9]。

收稿日期: 2014-04-20
基金项目:国家大学生创新性实验计划项目(1210504024);中央高校基本科研业务费专项(2013RW006);国家自然科学基金项目(61202305)
汪伟平,博士研究生. 研究方向:生物信息学、自然灾害学. E-mail: wangweiping0@gmail.com
通信作者:胡动刚,讲师. 研究方向:应用数学. E-mail: hudg@mail.hzau.edu.cn

Frank-Copula 函数的定义：

$$G_F(u,v)=-\frac{1}{\alpha}\ln\{1-\frac{(1-e^{-\alpha u})(1-e^{-\alpha v})}{1-e^{-\alpha}}\} \quad (1)$$

式中 α 为相关参数（$\alpha\neq 0$），当 $\alpha>0$ 时，随机变量 u 和 v 正相关；当 $\alpha\to 0$ 时，随机变量 u 和 v 趋于独立；当 $\alpha<0$ 时，随机变量 u 和 v 负相关。其中，相关参数 α 可以很好地衡量变量 u、v 之间的相关关系。同时，本文定义 2 个变量 X、Y 之间基于 Frank-Copula 函数的相关性为式(1)中的相关参数 α。实际上，由于其取值范围为$(-\infty,+\infty)$。数值范围较大且数值衡量变量之间的相关关系过于灵敏。但是，其 P 值却能很好地反映 Frank-Copula 拟合情况，进而反映变量之间的相关关系。因此，本文定义基于 Frank-Copula 函数的相关性测度为 $1-P$ 值，记为 $wcor(X,Y)\in[0,1]$，值越大表明变量间的相关性越好。

1.2 Copula 函数相关性测度优势

利用 R 语言编程，分别用 Person、Spearman、Kendall、Frank-Copula 4 种相关系数测度不同函数关系变量之间的相关性，结果如表 1 所示。

表 1 不同相关系数测度下变量 y 与 x 的关系

Table 1 The relationships between variables y and x's under different correlation coefficients measure

y 与 x 的关系 The relationships between y and x	相关系数 Correlation coefficients			
	Person	Spearman	Kendall	Frank-Copula
$y=x$	1.000	1.000	1.000	789.593
$y=\log(x)+x^2+\exp(x)$	0.610	1.000	1.000	789.593
$y=\tan(x)$	0.336	−0.025	−0.029	−0.190

从表 1 可以看出，Spearman、Kendall、Frank-Copula 测度法可以很好地衡量变量之间非线性关系和负线性关系。因此，从理论分析和数值模拟结果上均可以看出，基于 Frank-Copula 相关测度方法可以更好地衡量变量之间的相关关系。

2 基于 Copula 函数的权重基因共表达网络分析模型

2.1 构建基因共表达网络

网络可以被一个 $n\times n$ 维的邻接矩阵所确定，其中元素 $a_{ij}\in[0,1]$，表示网络中节点 i 和节点 j 之间的关联强度；共表达相似性 S_{ij} 用节点 i 和节点 j 之间相关性系数的绝对值表示[1]：

$$S_{ij}=|wcor(x_i,y_j)|$$

基于生物网络无标度理论，通过提高共表达相似性权重获得一个加权网络邻接矩阵[10]：

$$(a_{ij})=(s_{ij}^{\beta})$$

其中，β 值由无标度拓扑拟合指数（Scale-Free Topology Fitting Index）[11]得出。

2.2 发掘模块

权重基因共表达网络分析结果的好坏很大程度上由能否发掘出网络中正确的模块所决定。虽然模块挖掘是该模型今后急需讨论与工作重点关注之处，但是因文献[10]提供了便于计算的 R 语言包，所以本文仅利用其中的聚类方法发掘网络模块。

2.3 关联模块与样本性状信息

在模块发掘之后，本文将样本性状信息与模块相关联，从而找到与样本性状信息关系最密切的模块。假设发掘到 q 个模块，记为 $M_i(i=1,2,\ldots,q)$，那么对第 i 个模块内基因表达值和样本构成的矩阵 D_i 进行奇异值分解：

$$D_i=UD(V)'$$

定义　若记 U 的第一列为第 i 个模块的特征基因为 E_i [11]，则将样本性状信息表示为矩阵 t，定义第 i 个模块的样本性状信息重要性为

$$A_{e_i,t}=|wcor(t,E_i)|^{\beta}$$

其中 β 的与本文“2.1”中的取值一致。由此寻找所有模块中样本性状信息重要性值最大的模块，其与样本性状信息关系最为密切。进一步通过对候选基因的筛选、功能注释和基因本体论的深入探讨，可以得出模块内与样本性状信息最有关的信息。

3 小鼠体质量影响基因筛选

根据以上理论，用 R 语言编程实现基于 Copula 函数的相关性测度改进的权重基因共表达网络分析模型，对小鼠体质量基因数据（来源于 Langfelder 等[10]的研究）进行求解。

3.1 构建小鼠基因网络与发掘网络模块

构建网络并发现模块的结果如图 1 所示。

图 1 的上方表示 RNA 基因探针的聚类树状图；下方第 1 条色带为动态聚类结果图，第 2 条为动态聚类后合并距离较近的聚类结果图。其中，颜色表示模块类别，相同颜色的 RNA 探针属于同一模块。

3.2 关联模块与小鼠样本体质量信息

模块与外部形状关联结果如图 2 所示。

图 2 中最左列不同颜色表示不同模块，并用相应的颜色命名模块。将模块与性状之间的相关系数整理后如表 2 所示。由此，本文得到与体质量关联性最大的 magenta 模块，内含 107 个候选基因。

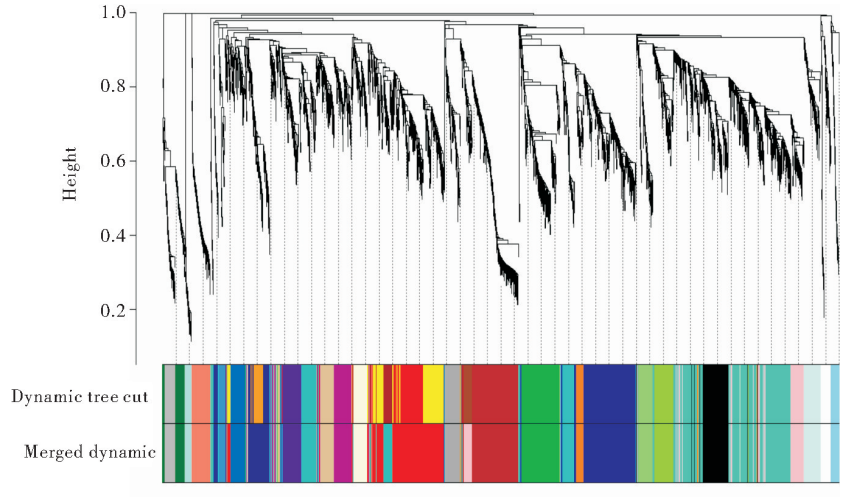

图 1　聚类图

Fig. 1　Cluster dendrogram

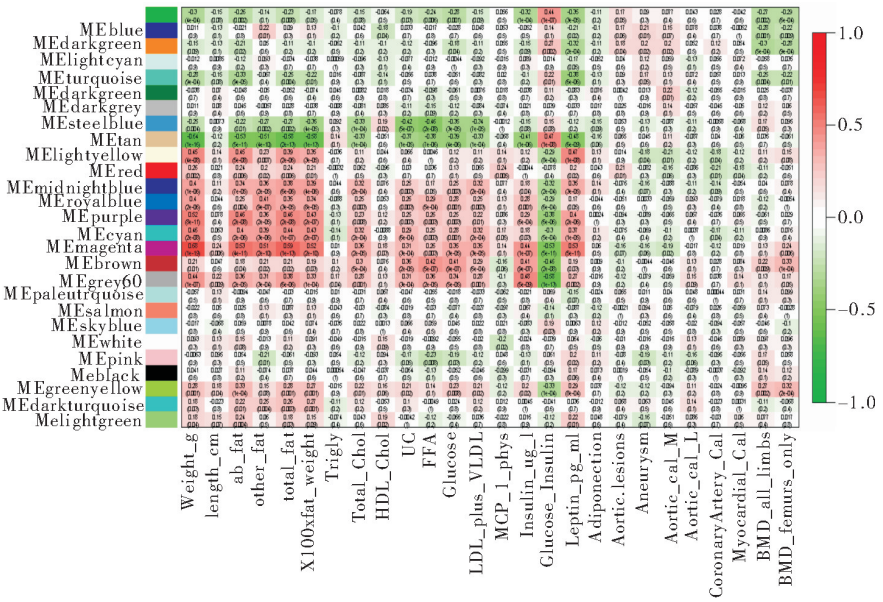

图 2　模块性状关联图

Fig. 2　Module-trait relationships

表 2　模块与体质量关联系数表

Table 2　Module-weight correlation coefficients

模块 Module type	关联系数 Correlation coefficients	模块 Module type	关联系数 Correlation coefficients	模块 Module type	关联系数 Correlation coefficients
Green	−0.302	Lightyellow	0.453	Paleturquoise	−0.057
Blue	0.011	Red	0.263	Salmon	−0.022
Darkorange	−0.148	Midnightblue	0.401	Skyblue	−0.017
Lightcyan	−0.011	Royalblue	0.399	White	0.093
Turquoise	−0.282	Purple	0.524	Pink	−0.006
Darkgreen	−0.038	Cyan	0.463	Black	0.041
Darkgrey	0.011	Magenta	0.682	Greenyellow	0.277
Steelblue	−0.245	Brown	0.212	Darkturquois	0.184
Tan	−0.638	Grey60	0.439	Lightgreen	0.180

4 生物机制与结果分析

4.1 小鼠体质量相关生物机制

小鼠体质量与发育相关，相同年龄段和生活环境的小鼠在体质量上存在差异的主要原因是肥胖。肥胖是体内脂肪堆积过多或分布异常的一种状态[12]，消除肥胖的主要生理过程为脂肪与葡萄糖的代谢。因为脂肪分解代谢反应在线粒体外，所以按文献[13]的方程分2步进行：第1步，甘油三酯和水在脂肪酶的作用下脂解生成甘油和脂肪酸；第2步，脂肪酸、ATP和脂酰CoA合成酶在Mg^{2+}的作用下转化成脂酰CoA、AMP和PPi[13]。

4.2 筛选结果基因类型一

降低血清甘油三脂、胆固醇和血清瘦素水平能促进能量消耗，加速脂肪氧化，提高高密度脂蛋白(HDL)水平，增加脂肪排泄，从而减轻动物的体质量[14]。在本文得到的结果中，基因*Cps*1参与甘油三酯的代谢过程。另有基因*Mogat*1等参与到甘油的代谢过程，基因*Gal*3*st*1、*Bucs*1、*Elovl*7和*Mogat*1参与脂质的代谢过程，*Elovl*7还参与了脂肪酸的合成和代谢过程，这些基因都是通过调节脂质代谢途径的物质来影响脂质代谢反应的平衡，从而影响脂质的代谢。

4.3 筛选结果基因类型二

环磷酸腺苷的蛋白激酶能使胞质内的甘油三酯脂肪酶磷酸化而活化，脂肪分解生成甘油和脂肪酸的速率增加[13]。本文的基因模块中，*Nrg*1、*Pdk*4、*F*7、*D*6*Ertd*245*e*、*AA*960558和*Stk*39等6个基因参与了蛋白激酶活性的调控。其中*Nrg*1正调控蛋白激酶，*F*7和*D*6*Ertd*245*e*则反映出负调控作用。

4.4 筛选结果基因类型三

*Fas*基因编码产物为分子质量45 ku的跨膜蛋白Fas蛋白与Fas配体结合后形成的Fas三聚体，使Fas胞质区死亡结构域(DD)相聚成簇，继而招募胞质内Fas相关死亡结构域蛋白(FADD)，并通过激活胱天蛋白酶(caspase)级联反应致使靶细胞走向凋亡[15]。在本文结果中与细胞凋亡有关的是*Ubd*基因和*Fsp*27基因。*Pparg*基因可以正向诱导激活细胞凋亡过程，*Msx*2基因和*Spp*1基因则负向调控细胞凋亡过程。由此猜测*Fas*等基因可以通过诱导细胞凋亡的途径来控制动物体脂的沉积。

4.5 筛选结果基因类型四

Vincent等[16]发现，肥胖同氧化应激相关，提出抗氧化的食品可以用来控制肥胖[17]。结果表明，*Fmo*3、*Cps*1、*Pparg*、*Apom*、*Gpx*4、*Cbr*3、*Cyp*2*g*1、*Aox*1、*Ephx*1、9330129*D*05*Rik*、*Npn*3、*Cyp*2*c*40等基因与氧化还原有关，涉及氧化还原、活化氧代谢、调节氧化还原酶活性等过程。

上述分析表明，通过基因功能注释分析，发现本文107个基因结果中有67个基因与导致小鼠体质量问题的因素直接相关，筛选结果有效率达到了62.6%。因此，本文改进的模型方法具有一定的科学性、有效性和可行性。同时，对小鼠体质量及肥胖相关基因的研究，可以为探索人体肥胖问题提供依据，并且对于寻找与人类肥胖有关的基因、更加科学地控制体质量等方面具有重要意义。

5 讨 论

本文从理论上和数值模拟上证明了基于Frank-Copula相关测度方法在相关性测度上的优势。在实际应用中，利用Frank-Copula函数改进权重基因共表达网络分析模型，通过求解影响小鼠体质量的基因数据并取得良好结果。因为目前对于基因功能的研究尚不完备，可能存在某些基因功能仍未发现的情况，所以本文所改进的模型筛选的结果可能会更好。但是，由于Copula函数参数估计方法相对于Person、Spearman等成熟的相关系数计算方法而言比较复杂，导致程序运行时间长达10 h甚至更长，从而影响模型的实用性。因此，本研究可以通过改进计算Copula函数参数估计方法来加快计算速度，提高模型的实用性。

参 考 文 献

[1] ZHANG B, HORVATH S. A general framework for weighted gene co-expression network analysis[J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): 1-45.

[2] MILLER J A, OLDHAM M C, GESCHWIND D H. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging[J]. The Journal of Neuroscience, 2008, 28(6): 1410-1420.

[3] WINDEN K D, KARSTEN S L, BRAGIN A, et al. A systems level, functional genomics analysis of chronic epilepsy[J]. PloS One, 2011, 6(6): e20763.

[4] PRESSON A, SOBEL E, PAPP J, et al. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome[J]. BMC Systems Biology, 2008, 2(1): 95.

[5] 籍艳丽. 基于 Copula 函数的秩相关和尾相关研究[J]. 经济问题，2009(5)：120-122.

[6] SKLAR M. Fonctions de répartition à n dimensions et leurs marges[J]. Publ Inst Stat Univ Paris，1959，8：229-231.

[7] 卢颖. Copula 理论在相关性分析中的应用及其多元扩展[D]. 天津：天津科技大学图书馆，2009.

[8] 朱新玲. 相关系数与 Copula 函数相关性比较研究[J]. 武汉科技大学学报：自然科学版，2009，32(6)：664-668.

[9] CHERUBINI U，LUCIANO E，VECCHIATO W. Copula methods in finance[M]. Chichester：John Wiley & Sons，2004.

[10] LANGFELDER P，HORVATH S. WGCNA：an R package for weighted correlation network analysis[J]. BMC Bioinformatics，2008，9(1)：559.

[11] HORVATH S. Weighted network analysis：applications in genomics and systems biology[M]. New York：Springer Book，2011.

[12] 李涛，籍保平，周峰，等. 红茶菌对饮食诱导肥胖小鼠体重控制的研究[J]. 食品科学，2009(11)：246-251.

[13] 吕志伟，马云霞，孙泽. 脂肪代谢的调节因素及运动对脂肪代谢的影响[J]. 伊犁师范学院学报：自然科学版，2010(1)：56-60.

[14] KAO Y H，CHANG H H，LEE M J，et al. Tea，obesity，and diabetes[J]. Molecular Nutrition and Food Research，2006，50(2)：188-210.

[15] 单安山，徐奇友. 动物脂肪代谢与调控[J]. 东北农业大学学报，2004，25(2)：129-134.

[16] VINCENT H K，INNES K E，VINCENT K R. Oxidative stress and potiential interventions to reduce oxidative stress in overweight and obesity[J]. Journal Compilation，2007，9(6)：813-839.

[17] VINCENT H K，TAYLOR A G. Biomarkers and potential mechanisms of obesity-induced oxidant stress in humans[J]. International Journal of Obesity，2006，30(3)：400-418.

Weighted gene co-expression network analysis based on Copula function

WANG Wei-ping[1,2]　BAI Ting[3]　HE Shuai[1,2]　LI Qian[4]　HU Dong-gang[5]

1. *State Key Laboratory of Earth Surface Processes and Resource Ecology，Beijing Normal University，Beijing* 100875，*China*；

2. *Academy of Disaster Reduction and Emergency Management，Ministry of Civil Affairs & Ministry of Education，Beijing* 100875，*China*；

3. *State Key Laboratory of Information Engineering in Surveying，Mapping and Remote Sensing，Wuhan University，Wuhan* 430079，*China*；

4. *College of Life Science and Technology，Huazhong Agricultural University，Wuhan* 430070，*China*；

5. *College of Science，Huazhong Agricultural University，Wuhan* 430070，*China*

Abstract　In order to improve the WGCNA model，we utilized Frank-Copula function to calculate correlation coefficient in this paper. Meanwhile，we applied the improved model to mine the mice genetic data，and discovered that there were 67 genes related to weight problems resulted from obesity in the maximum correlation coefficient module. The efficiency of the model was up to 62.6%，which justified the scientificity，effectiveness and feasibility of this model in functional genetic screening application.

Key words　weighted gene co-expression network analysis；correlation coefficient；Copula function；body-weight of mice；genetic screening

（责任编辑：边书京）