

## 基础与实验研究

## 急性高血糖对小鼠主动脉基因表达谱扰动的可视化研究

张心月, 俞梦越

## 摘要

目的: 采用数据挖掘方法, 基于 Matlab 平台对基因网络扰动可视化, 以直观显示短时高血糖状态下动脉血管转录组改变。

方法: 在美国国立卫生研究中心 (NCBI) GEO 数据库下载数据集。利用 Matlab 将数据转化为计算机可识别的结构体, 经过数据筛选, 获得短时高血糖后表达模式扰动最明显的基因谱。利用三种聚类算法分析, 基于 DAVID 进行基因本体学 (GO) 注释及富集分析, 把相关通路标定在 KEGG 通路中, 形成基因—表达谱系统分析。

结果: 经过对数据集的筛选、聚类将基因的变化模式归为 9 类, 在该模型中有效地反应出短时高血糖对动脉血管的急性早期效应。GO 富集分析显示, 在急性炎症反应、心肌重构、维持细胞内钙离子稳态、细胞周期调控、细胞趋化作用等方面的基因显著富集; 其中以与粘多糖、糖蛋白结构相关基因、脂肪分解代谢、肌原纤维组装相关基因显著富集。这些发现与以往研究的结论相吻合。K-均值聚类方法显示, 在高血糖环境下基因表达上调, 且不随血糖恢复正常表达的基因, 主要有参与细胞周期调控、心肌重构、维持细胞内钙离子稳态的基因。

结论: 利用数据挖掘方法, 实现急性高血糖对小鼠动脉基因表达谱波动模式的可视化描述, 并为糖尿病的“代谢记忆”机制提供新的解释, 即早期的高糖效应带来的动脉血管的不可逆的损伤, 是导致冠心病患者降糖治疗无效的原因。即短暂的高糖水平的暴露可在分子水平上起到长久的影响。

关键词 高血糖症; 内皮细胞; 基因表达谱

## Visualization Study on the Disturbance of Aorta Gene Expression Profile in Acute Hyperglycemia Mice Model

ZHANG Xin-yue, YU Meng-yue.

Department of Cardiology, Cardiovascular Institute and Fu Wai Hospital, CAMS and PUMC, Beijing (100037), China

Corresponding Author: YU Meng-yue, Email: yumy73@163.com

## Abstract

Objective: Based on the visualization function for gene network disturbance of Matlab platform, data mining method was used to directly observe transcriptional changes in aorta vessel at short-time hyperglycemia condition.

Methods: The information was down loaded from GEO database of NCBI. Using Matlab system to transfer the data set to a computer-readable structure, using data filter to obtain apparent gene expression disturbance profile after short-time hyperglycemia condition. Applying three clustering algorithms, based on DAVID platform to conduct gene ontology (GO) annotation and enrichment analysis in order to calibrate KEGG pathway and to form gene expression profile analysis.

Results: Via data set screening, the pattern of gene expression was divided into 9 clusters by special algorithms. GO analysis indicated that obvious gene enrichments were found in acute inflammation reaction gene, myocardium remodeling gene, stabilizing intracellular calcium gene, cell cycle regulation gene, chemotactic effect gene; especially in mucopolysaccharide gene, glycoprotein structure related gene, fat catabolism gene and myofibril related gene. The above findings were identical to previous study. K-means clustering method presented that in hyperglycemia condition, up-regulated genes didn't return to normal level when blood glucose back to normal which mainly including cell cycle regulation gene, myocardium remodeling gene and stabilizing intracellular calcium gene.

基金项目: 国家自然科学基金委员会面上项目 (81670415)

作者单位: 100037 北京市, 北京协和医学院 中国医学科学院 国家心血管病中心 阜外医院 冠心病诊治中心

作者简介: 张心月 硕士研究生 主要从事冠心病基础研究 Email: 18800161600@163.com 通讯作者: 俞梦越 Email: yumy73@163.com

中图分类号: R541 文献标识码: A 文章编号: 1000-3614 (2017) 09-0924-06 doi: 10.3969/j.issn.1000-3614.2017.09.023

**Conclusion:** Our work provided a new explanation of diabetes metabolic memory; short-term hyperglycemia caused arterial damage was irreversible which incurred inefficient hypoglycemic therapy in coronary artery disease patients.

**Key word** Hyperglycemia; Endothelial cells; Gene expression profile

(Chinese Circulation Journal, 2017;32:924.)

糖尿病是冠状动脉疾病(CAD)的独立危险因素,其中动脉粥样硬化在糖尿病患者死亡原因中约占80%<sup>[1,2]</sup>。糖尿病可通过多种途径促进动脉粥样硬化的发生,如高血糖、肥胖、血脂紊乱、高血压、胰岛素抵抗等;其中机体处于长期慢性高糖环境,是糖尿病患者发生动脉粥样硬化的主要因素<sup>[3,4]</sup>。高血糖可通过多条途径启动动脉粥样硬化的发生、并加速其进展,其中内皮细胞功能受损是动脉粥样硬化和糖尿病大血管病变的早期病理生理变化。

DNA 微阵列(基因芯片)是高通量研究中重要的方法之一,能对大量的基因表达谱进行同步、快速检测,提供上万条基因的表达谱。目前公共数据库(如 GEO)中基因芯片表达谱数据与日俱增,但海量的数据却未得到充分的提取与深入的挖掘。以往的研究更多的局限于传统的研究、统计手段,但芯片数据的特点是维数高、具有异质性、网络性,传统的统计分析方法不再适用。数据挖掘是在大数据背景下应运而生的一种将数据转换为有用信息的方法<sup>[5]</sup>。本文结合计算生物学技术,采集公共数据库中已发表的 Affymetrix 寡核苷酸微阵列原始数据 GDS4016,对数据集进行数据筛选、差异表达基因筛选、聚类分析、基因本体学注释、通路富集分析,从而更加直观地显示短时高血糖对血管内皮细胞的基因扰动,以利于复杂信息进一步整合、挖掘。

## 1 材料与方法

数据集来自美国国立卫生研究中心(NCBI)的 GEO 数据库,编号为 GDS4016,平台号为 GPL1261(Affymetrix 公司),下载地址为: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15401>。该芯片主要研究内容为瞬时高血糖过程中及之后,小鼠主动脉组织细胞基因表达谱改变。实验检测时间点为第 0 天(血糖 5 mmol/L)、第 2、4、7 天(血糖 >25 mmol/L)、第 11、26 天(血糖 5 mmol/L)。

数据预处理:①弱信号处理:芯片上存在很多弱信号点,这些点的信号强度虽然很弱,但不一定

是噪声点,有可能是一些十分重要的基因,因此不能武断地把它们全部删去。目前对于如何把弱信号点从背景或噪声中分离出来仍无全面有效的方法。我们通过背景、空白点、阴性对照点确定弱信号的阈值。低于该值的信号点被滤除,高于该值的信号点进入后续的数据分析。②数据的对数转换:对数转换能够提供从生物学角度易于解释的数据,使数据的分布满足近似正态分布,以便后续的数据挖掘方法的使用。为确保聚类的有效性,使芯片数据偏向于正态分布,对数据集进行标准对数转换。③数据筛选:基因表达谱数据集很大,大部分组织只有大约 10000~15000 个基因会产生表达<sup>[6]</sup>,并且很多基因在实验中没有表现出我们感兴趣的变化。为了简化搜寻兴趣基因的过程,需要缩减数据集的数量到一个亚集中,去掉不感兴趣的基因。因而我们采用低熵筛选,过滤掉表达量波动熵值过低(基本熵值属于 10% 以下)的基因。

描述表达谱扰动的算法:无监督学习(unsupervised analysis)即没有事先定义的向量集或类别集,使用递归的分割方法来分类,把拥有相似特征的数据归入相同的类。包括层级聚类、K-均值聚类、自组织映射、主成份分析等。①K-均值聚类(K-means Clustering):K 均值聚类是无监督分类的一种基本方法<sup>[7]</sup>。在 K-means 算法运行前须指定聚类的数目 K 和迭代次数,并指定 K 个初始质心。初始质心数目的选取,一种是随机选择,另一种是使用其他聚类方法得到的类平均向量<sup>[8]</sup>。②主成份分析(principle component analysis, PCA):是一种数学降维方法,找出几个综合变量来代替原来众多的变量,使这些综合变量能尽可能地代表原来变量的信息量,并且彼此之间相互独立。这些新变量能够代表原始数据的能力由它们所能解释的变异的比例衡量。那些远离远点的基因,是表达量变化最明显的基因,在得分图中选取离原点较远的点,对对应的基因进行标记<sup>[9]</sup>。③自组织映射(Self-organizing Maps, SOMs):是一种无监督神经网络,主要用于对输入向量的区域分类。它克服了 K-means 聚类的一些缺点,如对噪声稳定,不依赖与数据分布的形状,提供大数据集内相似性关系的综合分析。联合

PCA 和 SOM 用于基因数据的聚类分析,比单一使用 SOM 的聚类分析有更高的分类正确率及较为清晰的分类边界<sup>[10]</sup>。

研究分析平台:基于 Windows7 (Microsoft, 美国)的 Matlab 2013a (MathWorks, 美国)的生物信息工具包。

## 2 结果

### 2.1 数据预处理

数据转换(图 1):GDS4016 数据集中每个时间点包含三个生物学重复的样本,对其求平均值,并进行对数转换后的数据在各个时间点的散点图。

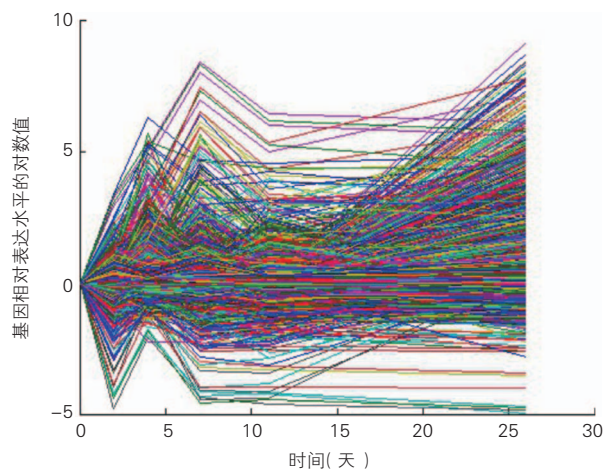


图 1 数据转换后 45 101 个芯片数据的折线图

基因筛选:首先根据数据挖掘思路,对数据进行去空值清洗,然后筛选掉复制次数波动小于 2.5 次的基因,及变化水平 2 倍以下的基因;利用 MATLAB 中的低熵筛选,滤除表达量波动熵值过低(基本熵值属于 10% 以下)的基因,获得瞬时高血糖后表达模式扰动最明显的基因表达谱。

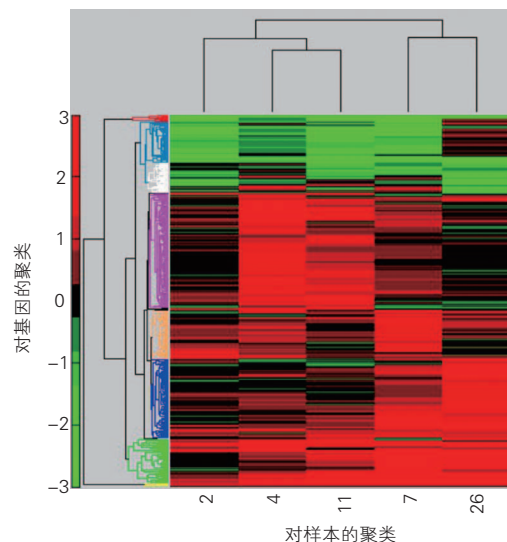
这样从 45 101 条序列中挑选出 686 条有意义波动的基因,作为感兴趣基因。

### 2.2 K 均值聚类算法

对筛选后的数据集,通过分层聚类结果构建热点图、系统树图(图 2),以直观的了解其表达情况。

利用层级聚类发现结果簇的数目可大致分为 9 种情况,设定 K 均值聚类的分类数 K 为 9,初始质心的选择依靠计算随机生成,使用欧氏距离,并设置最大迭代次数为 100 次。K-means 聚类算法聚类

图如图 3 所示。可看出基因的表达模式被归入 9 簇,随时间序列的变化,基因的表达具有明显的上调或下调。其聚类的质心(图 4)绘出这 686 条基因的表达轮廓。



注:图中横坐标表示对样本的聚类,一列对应一个样本,样本间距离越近表示基因表达越类似。纵坐标表示对基因的聚类,基因之间距离越近则表达水平越接近。色阶表示基因表达丰度,绿色表示基因表达下调,红色表示表达上调

图 2 双向分层聚类结果

### 2.3 主成份分析

PCA 是对大数据集降维的重要工具,亦可在噪声数据中发现信号。首先对 686 条基因进行主成份分析,图 5 显示出第一二个主成份的散点图,表示出有两个不同的区域,因为筛选功能已经把低变化及低信息量的数据去除了,而这些点应该出现在散点的中心。6 个主成分的特征值及它们的贡献率见表 1。从表中可以发现,前 2 个特征值的贡献率达到近 90%,包含了原始数据集的主要信息。

主成份分析中,发现异常的离群点,说明这些基因的表达模式与其他大多数的基因并不相同。利用 DAVID 分析,发现这些基因在细胞组成上,主要在肌原纤维组成富集。其生物学作用主要与肌肉组织发育分化相关,如心肌肌动蛋白  $\alpha 1$  (Actin, Alpha, Cardiac Muscle 1, ACTC1)、锚蛋白重复域 1 (Ankyrin Repeat Domain 1, ANKRD1)和心肌钙蛋白 T (Troponin T2, Cardiac Type, TNNT2);另外,与细胞骨架重塑的基因也有显著富集,如肌球蛋白重链 6 (Myosin Heavy Chain 6, MYH6)、肌联蛋白 (Titin, TTN)等。



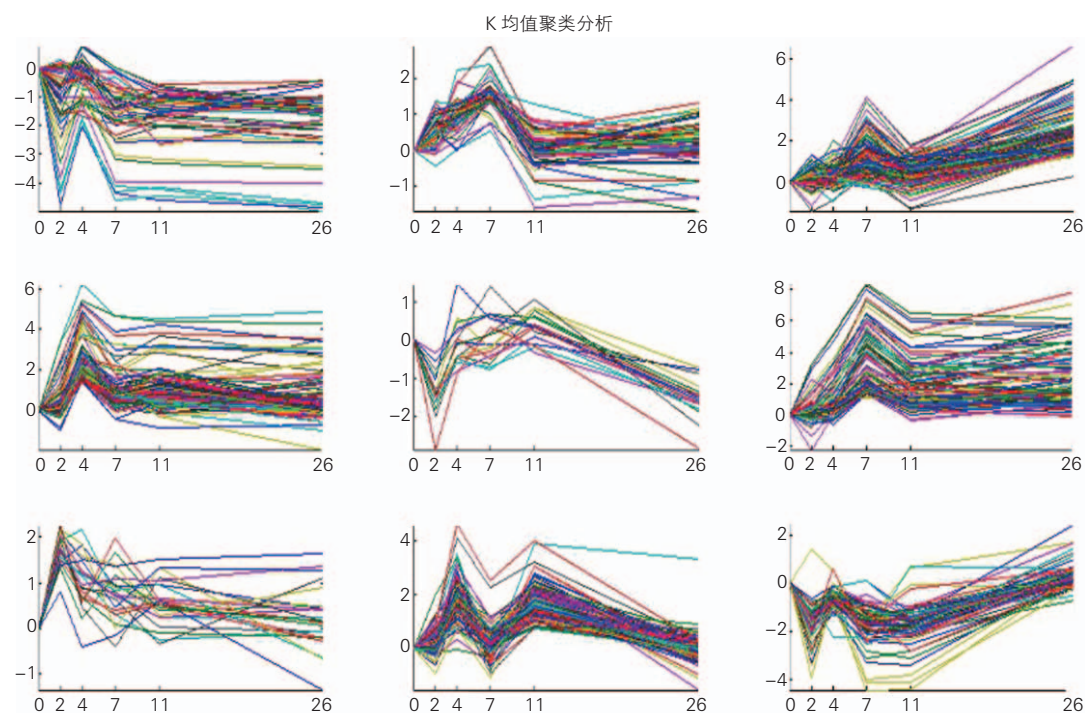


图3 K-means 聚类算法将 686 个数据归为 9 簇

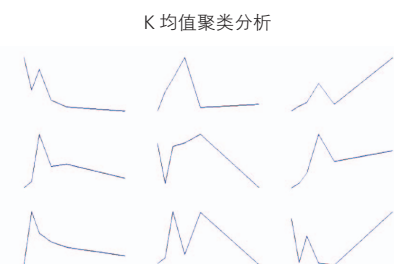


图4 K-means 算法基因随时间序列的表达轮廓

表1 主成分的特征值与贡献率

| 主分量 | 特征值      | 贡献率(%)  | 累计贡献率(%) |
|-----|----------|---------|----------|
| 1   | 7.336044 | 69.5343 | 69.5343  |
| 2   | 1.866844 | 17.6948 | 87.2291  |
| 3   | 0.604724 | 5.7319  | 92.961   |
| 4   | 0.398636 | 3.7785  | 96.7394  |
| 5   | 0.343998 | 3.2606  | 100      |
| 6   | 0        | 0       | 100      |

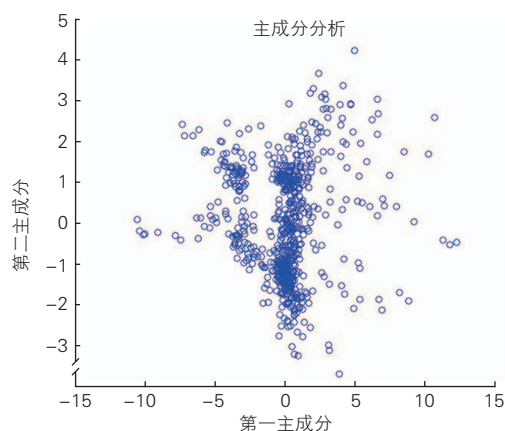


图5 主成份分析

## 2.4 自组织映射(图6)

把 PCA 过程中得到的前两个特征输入 SOM 作为其输入变量,用前两个主成份构建 SOM,用系统设定的参数训练网络,这种方法减少了在训练过程中关联度不大的基因的影响,能够有效地提高网络训练速度及聚类准确率。总共聚类为 16 类,不同的颜色标识出不同的类别,16 个红色的点代表聚类的中心。

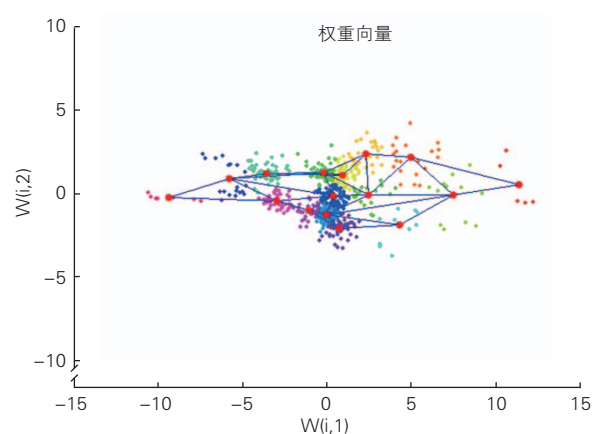


图6 针对主成分分析结果的自组织映射聚类

## 2.5 基因本体学与通路分析

根据三种聚类算法得到有相似表达谱的基因,这意味着有相同表达模式的基因(在同一个聚类簇中)在同一的处理条件下表现为共同上调或者下调,

这样我们给出一种假设,即这些基因异同执行一个特定的生物学功能。对不同的 GO 术语和 KEGG 通路进行的富集分析,把功能相似的基因聚在一类,并分析它们在整个基因组表达背景下的富集情况,以便更加直接地了解基因所代表的生物学功能。

运行 DAVID,利用统计学分析(Fisher 精确概率检验),计算出  $P$  值, $P$  值小于设置(0.05),表明基因有显著富集。对 Fisher 精确概率检验进行修饰的 EASE 得分,得分越高,富集效果越好。对筛选出的基因做富集分析,利用 DAVID 中的功能注释富集,对 686 个基因进行分析,可以看出这些筛选后的基因主要分布在肌原纤维,并在瘢痕修复、免疫反应有显著富集,这与之前的研究相一致<sup>[11]</sup>。通路富集分析显示这些基因在肥厚型心肌病通路上显著富集。下面具体分析 K-means 分类下 9 类基因中我们感兴趣的变化趋势的基因的生物学功能。

第一类,基因表达变化基本分布在 0 以下,表明相较第 0 天(高糖作用前小鼠主动脉组织)基因表达下调;第 2 天低血糖的窗口期,基因表达出现明显下调;第 4 天随着血糖水平恢复,基因表达水平随之恢复;第 7 天,高血糖的峰值时,基因再次下调,且与第 2 天下调幅度接近。但第 7 天之后,血糖逐渐恢复正常,并在正常水平持续了近两周,下调的基因未与之相应地调回原来的水平。基因本体学分析结果显示,该类基因主要与急性期免疫应答相关,如血清淀粉样蛋白(Serum Amyloid A1, SAA1)、丝氨酸蛋白酶抑制因子 1(Serpin Family A Member 1, SERPINA1)。第九类的表达趋势与第一类相类似,生物学作用方面主要与细胞迁移相关,如整合素  $\alpha 4$  (Integrin Subunit Alpha 4, ITGA4)、整合素  $\beta 2$  (Integrin Subunit Beta 2, ITGB2)、FC 段  $\gamma$  受体 3 (Fc Fragment Of IgG Receptor III, FCGR3) 以及 S100 钙结合蛋白 A9 (S100 Calcium Binding Protein A9, S100A9);并与胆固醇代谢相关,如载脂蛋白 A1 (Apolipoprotein A1, APOA1)、瘦素(Leptin, LEP)和 SAA1。

第二类中,基因转录的大致趋势呈现先增高后降低,第 7 天(血糖最高值)基因明显上调达到峰值,随后随着血糖水平逐渐恢复至正常(第 11 天),基因的转录水平也逐渐下降至第 0 天的水平,并在其后 15 天维持该表达水平。富集分析结果显示,该类基因主要表达糖蛋白,并主要聚集在细胞外区域,如 CD99L2、FRAS1、组织蛋白酶 H (Cathepsin H, CTSH) 等。钙化相关基因显著富集,如骨成型

蛋白受体 2 (Bone Morphogenetic Protein Receptor Type 2, BMPR2)、WNT 抑制因子 1 (WNT Inhibitory Factor 1, WIF1)、骨调蛋白 (Osteomodulin, OMD);以及血管细胞粘附分子 1 (Vascular Cell Adhesion Molecule 1, VCAM1)、基质金属蛋白酶 3 (Matrix Metalloproteinases 3, MMP3)、血管性血友病因子 (Von Willebrand Factor, VWF)、载脂蛋白 D (Apolipoprotein D, APOD)、JUN 和肌球蛋白重链 6 (Myosin Heavy Chain 6, MYH6);并与急性免疫反应密切相关。这与原始研究结果相一致<sup>[12]</sup>。除此之外,研究还发现对氧磷酶 1 (Paraoxonase 1, PON1)、TNN1、肌原调节蛋白 2 (Myozenin 2, MYOZ2) 也包括在该类基因当中。

第三类与第六类的表达情况相类似,呈现先增高后降低的趋势,第 7 天达到峰值。与第二类的区别在于,第 11 天以后表达水平仍然高于第 0 天的表达水平,表明基因的转录水平未得到逆转。第四类与这两类的区别在于,基因上调的最高值落在第四天。富集分析的结果显示,除了原始研究中发现的表达未逆转的基因,如血管生成素 (Angiogenin, ANG)、花生四烯酸-15-脂加氧酶 (Arachidonate 15-Lipoxygenase, ALOX15)、载脂蛋白 B-mRNA 编辑酶复合物 1 (Apolipoprotein BmRNA Editing Enzyme Catalytic Subunit 1, APOBEC1)、FOS、肌酸激酶 (Creatine Kinase, M-Type, CKM)、早期生长应答蛋白 3 (Early Growth Response 3, EGR3) 之外,我们发现这类基因还包括与肌原纤维组装、细胞周期调控、维持细胞内钙离子稳态和蛋白锚定相关的基因。

第八类中,基因转录水平形成先升后降、再升再降的趋势,并在第 4、11 天,血糖分别从低糖、高糖恢复至正常的时间达到峰值,第 26 天恢复至第 0 天的水平。说明该类基因的表达滞后于血糖的变化。对这一类基因做富集分析,发现在细胞功能上,与趋化作用相关的基因显著富集,包括 FC 段  $\gamma$  受体 2 (Fc Fragment Of IgG Receptor II, FCGR2)、S100 钙结合蛋白 A8 (S100 Calcium Binding Protein A8, S100A8) 等。

第五类和第七类的基因数较少,不做富集分析,单独分析每一个基因,未从中发现有意义的点。

综上所述,该研究发现了除原始分析结果外其他受调控基因,如 PON1、FRAS1、CTSH、TNN1、MYOZ2 等。其中表达未能逆转的基因还包括与肌原

纤维组装、细胞周期调控、离子转运、维持细胞内钙离子稳态和蛋白锚定相关的基因。对受调控基因的变化趋势进行进一步细分,有利于深入理解疾病变化过程的具体机制,促进进一步深入研究。

### 3 讨论

目前公共数据库中基因芯片表达谱数据与日俱增,但海量的数据却未得到充分的提取与深入的挖掘。采集公共数据库中已发表的 Affymetrix 寡核苷酸微阵列原始数据 GDS4016,利用数据挖掘的方法,从新的角度对已有的芯片进行崭新的研究。基于 Matlab 中 Mathwork 生物信息工具包将数据转化为计算机可识别的结构体,经过数据转化使数据在同一水平上可比较,之后运用模式识别算法剔除表达背景噪声,获得短时高血糖后表达模式扰动最明显的基因谱。利用 K-means 算法、主成份分析及自组织图对筛选后的基因进行聚类。实现急性高血糖对小鼠动脉基因表达谱波动模式的可视化描述。进而基于 DAVID 进行 GO 注释及富集分析,把相关通路标定在 KEGG 通路中,形成基因——表达谱系统分析。解释了为什么降糖处理不能降低 CAD 的发生风险,即短暂的高糖水平的暴露可在分子水平上起到长久的影响。目前对高血糖症的治疗指南,没有对及时降糖给予足够的强调,但是实验与临床研究均表明,持续的动脉损伤会加速动脉粥样硬化及心脏病的进程,早期的低糖或高糖效应带来的动脉血管的不可逆的损伤,是导致冠心病患者降糖治疗无效的原因。

本研究虽从研究方法上为高血糖相关的冠心病研究提供了新的途径,但仅从转录组水平进行数据分析就得出肯定结论尚欠缺说服力,最终还需得到细胞或动物模型的进一步验证。其次,本研究数据来自美国数据库,基于遗传异质性和人种差异,未

来基于中国人的数据库分析更有指导意义。

随着大规模测序的发展,云存储技术的运用,在未来,数据每日的更新量会令人瞠目结舌。如何有效的利用这些资源,把它们转化为我们能够利用的信息,必须借助新的算法的开发。数据挖掘的方法能够帮助我们在海量的信息当中发现关联,形成能够被利用的知识。

### 参考文献

- [1] Martín-Timón I, Sevillano-Collantes C, Segura-Galindo A, et al. Type 2 diabetes and cardiovascular disease: Have all risk factors the same strength? *World J Diabetes*, 2014, 5: 444-470.
- [2] Milicevic Z, Raz I, Beattie SD, et al. Natural history of cardiovascular disease in patients with diabetes: role of hyperglycemia. *Diabetes Care*, 2008, 31 Suppl 2(Supplement 2): S155-S160.
- [3] Aronson D, Rayfield EJ. How hyperglycemia promotes atherosclerosis: molecular mechanisms. *Cardiovasc Diabetol*, 2002, 1: 1.
- [4] Nagareddy PR, Murphy AJ, Stirzaker RA, et al. Hyperglycemia promotes myelopoiesis and impairs the resolution of atherosclerosis. *Cell Metab*, 2013, 17: 695-708.
- [5] Wu F-X, Li M, Ruan J, et al. Systems biology approaches to mining high throughput biological data. *Bio Med Research International*, 2015, 2015: 504362-504362.
- [6] Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA*, 2002, 99: 4465-4470.
- [7] Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *Int J Comput Assist Radiol Surg*, 2016, 11: 2033-2047.
- [8] Fernandez EA, Balzarini M. Improving cluster visualization in self-organizing maps: application in gene expression data analysis. *Comput Biol Med*, 2007, 37: 1677-1689.
- [9] 蔡斌, 江华. 急性心肌梗死早期基因表达和代谢调控网络扰动的可视化研究. *中华急诊医学杂志*, 2013, 22: 591-596.
- [10] 程国建, 安瑶. 基于 PCA 的 SOM 网络在基因数据聚类分析中的应用. *软件导刊*, 2013, 12: 127-130.
- [11] 冯新星, 陈燕燕. 糖尿病心肌病的研究进展. *中国循环杂志*, 2015, 30: 87-89.
- [12] 陈丽莉, 范国治, 韩蕊, 等. 二甲双胍降低 2 型糖尿病大鼠主动脉磷酸化丝裂原活化蛋白激酶的蛋白表达. *中国循环杂志*, 2015, 30: 487-491.

(收稿日期:2016-11-26)

(编辑: 汪碧蓉)