

基于表达及网络拓扑结构挖掘动脉粥样硬化风险疾病基因*

王 宏 曲晓莉 赵 研 张 静 陈丽娜**

(哈尔滨医科大学生物信息科学与技术学院, 哈尔滨 150086)

摘要 基于功能基因组信息、网络拓扑结构信息整合分析方法, 利用基因表达谱数据和蛋白质互作数据挖掘动脉粥样硬化(AS)风险疾病基因, 为从基因组层面研究动脉粥样硬化提供了新的视角. 经过差异表达分析, 支持向量机(SVM)的机器学习方法双重筛选, 可以鉴别出可信度水平较高的风险疾病基因, 对于研究动脉粥样硬化疾病基因在网络中的拓扑性质, 建立基因与疾病发生发展过程的联系, 提供了新的思路. 得到了巨噬细胞样本中 59 个风险疾病基因, 泡沫细胞中 61 个风险疾病基因. 这些风险基因与已知疾病基因共享大部分动脉粥样硬化病变相关生物学过程及信号通路. 并应用到对其他复杂疾病致病机理的研究中.

关键词 动脉粥样硬化, 差异表达, 网络拓扑特征, 支持向量机

学科分类号 R318.04, Q78

DOI: 10.3724/SP.J.1206.2009.00687

人类复杂疾病是由多个基因共同作用导致的, 生物医学的一个关键目标是寻找风险疾病基因并阐述其在复杂疾病致病过程中所起的重要作用^[1]. 目前, 许多方法被应用到风险疾病基因预测研究中, 如基因组关联分析、疾病基因与非疾病基因间序列与功能差异分析等.

随着基因芯片技术的快速发展, 基于全基因组表达水平的疾病发病机理分析成为可能. 通过对海量数据的整理、分析, 能够了解基因表达变化与病理特征的对应关系, 从而进行疾病诊断、类型预测以及疾病网络研究^[2].

蛋白质互作是细胞内生物学过程的基础, 细胞内各种生物分子相互联系, 行使功能^[3]. 通过分析蛋白质互作网络拓扑性质, 能够预测基因功能, 阐述复杂疾病病理生理作用机制^[4]. 蛋白质互作网络研究, 已经成为系统生物学和生物信息学的热点.

但是, 单一从网络分析或基因表达分析, 不能全面准确阐明基因与疾病之间的复杂联系. 因此, 本文提出一个创新的整合分析方法, 从功能基因组信息与网络拓扑结构整合角度挖掘动脉粥样硬化风险疾病基因, 揭示动脉粥样硬化发生发展机制. 本方法对疾病基因网络属性以及相关生物学功能、信

号转导通路的深入分析, 为动脉粥样硬化风险疾病基因的挖掘及致病机理的研究提供了新的思路.

1 材料与方法

1.1 数据收集

人类蛋白质互作数据来自于 HPRD 数据库(版本 7), 去掉蛋白质自身互作, 得到包含 9 013 个蛋白质, 34 032 个互作的蛋白质互作网络^[5]. 动脉粥样硬化表达谱数据 GSE9874 来自于 GEO 数据库^[6-7], 包含 22 283 个探针, 60 个样本, 其中 15 个动脉粥样硬化巨噬细胞样本, 15 个非疾病巨噬细胞样本, 15 个动脉粥样硬化泡沫细胞样本, 15 个非疾病泡沫细胞样本. 从 OMIM 数据库中得到 124 个动脉粥样硬化已知致病基因^[8].

* 哈尔滨医科大学研究生创新基金资助项目(HCXS2010006), 黑龙江省自然科学基金资助项目(D2007-48), 国家自然科学基金资助项目(30571034)和哈尔滨医科大学大学生创新基金(HRBCX2008-02)资助项目.

** 通讯联系人.

Tel: 0451-86674768, E-mail: chenlina_2004@yahoo.com.cn

收稿日期: 2010-03-25, 接受日期: 2010-05-17

基于基因表达与蛋白质网络拓扑结构整合双重筛选, 挖掘动脉粥样硬化风险疾病基因的具体流程如下(图 1)。

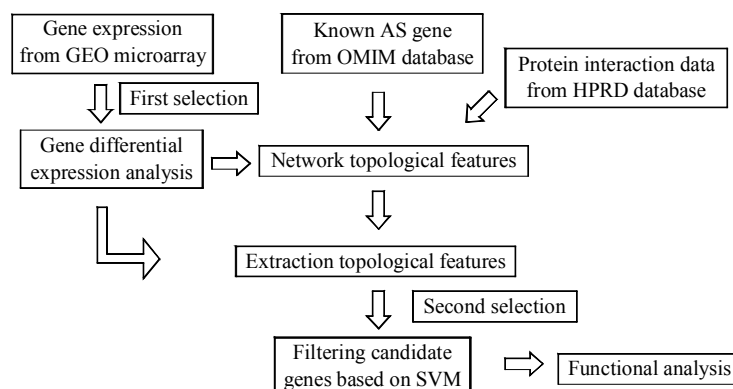


Fig. 1 Work flow

1.2 差异表达基因筛选

首先, 对于基因表达谱数据进行预处理, 对数转换、补缺失值及合并重复基因。其次, 利用 t 检验挖掘动脉粥样硬化巨噬细胞样本表达谱及泡沫细胞样本表达谱中的差异表达基因。我们将显著性 P 值小于 0.05 的基因定义为在动脉粥样硬化疾病状态下显著差异表达的基因, 选择这些差异表达基因为候选基因(第一步筛选结果)。

1.3 网络拓扑性质

本文分析了疾病基因与非疾病基因在蛋白质网络中的三个拓扑测度(度、聚类系数、最短路径)。度提供了节点在网络中连接程度信息。如果一个节点的邻居较多, 则它在网络中较为重要, 我们称之为 Hub 节点^[9]。聚类系数刻画了节点周围邻居间联系的紧密程度。最短路径是指从一个节点到另一个节点所经过的最少的连接数。本文使用与已知致病基因的平均最短路径来描述基因到已知致病基因的距离。

1.4 支持向量机筛选风险基因

支持向量机是数据挖掘技术中有效地用于分类的机器学习方法, 广泛应用于生物信息学研究领域。首先确定训练集, 本文选取的阳性训练集为 OMIM 数据库中的已知动脉粥样硬化致病基因集, 阴性训练集为蛋白质互作网络中去除已知致病基因和差异表达基因后所剩余的基因集合(也称作非致病基因)。其次确定检验集, 本文选取动脉粥样硬化巨噬细胞样本中的差异表达基因、动脉粥样硬化

泡沫细胞样本中的差异表达基因作为检验集。为了更好地挑选阳性训练集和阴性训练集, 我们采用十倍交叉证实的方法来确定训练集。最后分别利用度、聚类系数及最短路径三种网络拓扑特征作为支持向量机的分类特征, 对检验集进行筛选, 得到动脉粥样硬化风险疾病基因。

1.5 风险疾病基因功能分析

应用 Gene Ontology^[10]和 Biocarta^[11]数据库, 对利用支持向量机预测得到的风险疾病基因与已知疾病基因进行分子功能、细胞组分和信号转导通路分析。探讨已知疾病基因和风险疾病基因在动脉粥样硬化发生发展过程中的作用。

2 结 果

2.1 差异表达基因

GEO 数据库中的动脉粥样硬化表达谱数据 GSE9874, 包含 60 个样本(30 个疾病样本, 30 个对照组样本)。利用 t 检验获得芯片数据中的差异表达基因, 设定显著性阈为 0.05, 结果得到动脉粥样硬化巨噬细胞样本表达谱差异表达基因 135 个, 动脉粥样硬化泡沫细胞样本表达谱差异表达基因 212 个。将这些差异表达基因定义为候选基因(第一步筛选结果)。

2.2 网络拓扑性质分析

分别计算网络中已知致病基因与非疾病基因的程度、聚类系数和平均最短路径, 比较分析结果如图 2, 3, 4。

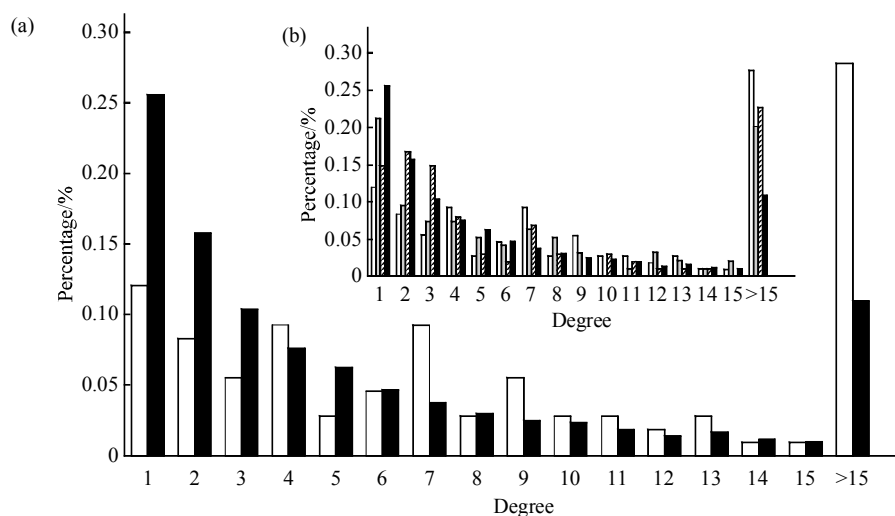


Fig. 2 Degree of disease gene and non-disease gene

□: Disease gene; □: Macrophages sample; ▨: Foam cell sample; ■: Non-disease gene.

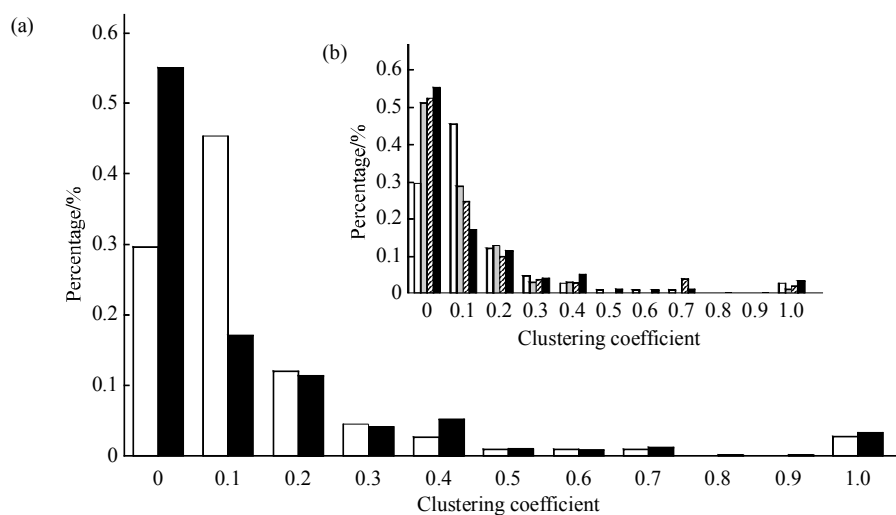


Fig. 3 Clustering coefficient of disease gene and non-disease gene

□: Disease gene; □: Macrophages sample; ▨: Foam cell sample; ■: Non-disease gene.

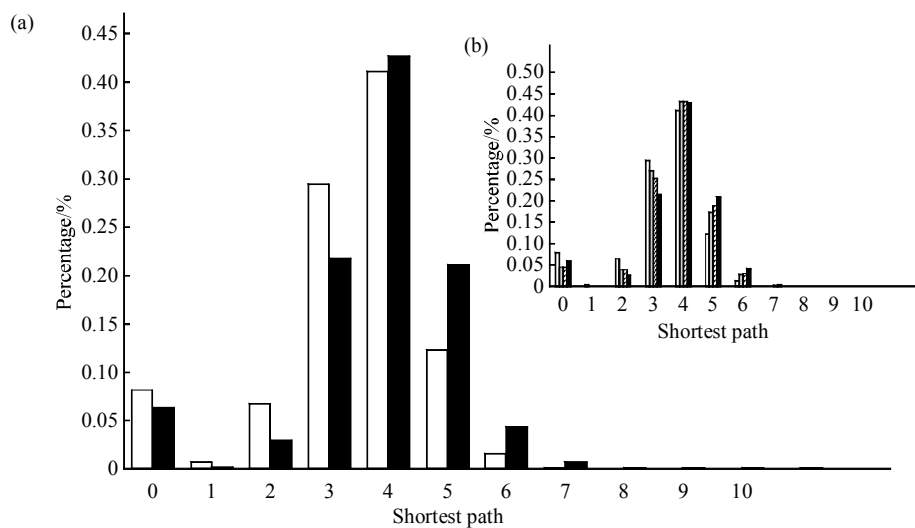


Fig. 4 Shortest path of disease gene and non-disease gene

□: Disease gene; □: Macrophages sample; ▨: Foam cell sample; ■: Non-disease gene.

从度分布图(图 2)中可以发现, 已知致病基因与非疾病基因的度分布存在着明显差异. 在网络中非疾病基因一般具有较小的连接, 度分布集中在 6 以下(70%). 而对于已知致病基因, 当度大于 6 时, 基因所占的比例(60%)明显高于非疾病基因所占的比例, 当度大于 15 时尤为明显. 可见动脉粥样硬化致病基因多分布在网络的 hub 节点上.

从聚类系数分布图(图 3)中可以发现, 已知致病基因的聚类系数主要分布在 0.1~0.2 的范围内, 要显著高于非疾病基因(55%: 28%), 而非疾病基因的聚类系数主要分布在 0 处(55%). 可见动脉粥样硬化致病基因倾向于紧密连接.

从最短路径分布图(图 4)中可以看出, 最短路径 0~3 范围内, 已知致病基因比例(约 45%)要高于非致病基因(约 30%), 而 4~7 范围内, 已知致病基因比例(约 50%)低于非疾病基因(约 68%). 通过分析基因到已知致病基因间的最短路径, 可见已知致病基因之间倾向于 1、2、3 次互作, 具有十分紧密的连接性.

通过以上分析, 我们发现动脉粥样硬化已知疾病基因通常具有较高的连接度, 分布在网络中的 hub 节点上. 而这些已知致病基因之间通常倾向于紧密地连接, 并且与非致病基因存在着显著差异.

因此通过对动脉粥样硬化致病基因在网络中的拓扑特征分析, 我们发现度、聚类系数、最短路径可以用来作为利用支持向量机区分致病基因与非致病基因的分类特征, 进而对检验集样本进行预测筛选.

同时, 我们还比较了巨噬细胞样本中差异表达基因、泡沫细胞样本中差异表达基因与已知致病基因以及非疾病基因的度、聚类系数和平均最短路径(图 2b~4b), 从图中可见巨噬细胞样本差异表达基因和泡沫细胞样本差异表达基因, 与已知致病基因的拓扑性质大致相同, 而非疾病基因存在着明显差异.

2.3 利用支持向量机筛选风险疾病基因

基于三个网络拓扑特征: 度、聚类系数及最短路径, 利用支持向量机对阳性训练集和阴性训练集进行训练, 随后对检验集进行分别筛选, 选取三个分类特征筛选结果的交集, 最终得到动脉粥样硬化巨噬细胞样本中风险疾病基因 59 个, 动脉粥样硬化泡沫细胞样本中风险疾病基因 61 个.

2.4 风险疾病基因功能分析

我们对动脉粥样硬化已知致病基因, 与支持向量机筛选得到的风险疾病基因进行功能富集分析, 利用 Cytoscape 软件中的 BINGO 插件, 基于超几何检验(显著性水平为 0.05), 分析风险疾病基因与已知致病基因在分子功能(MF), 细胞组分(CC)上的富集情况. 统计结果显示, 已知致病基因与动脉粥样硬化巨噬细胞样本、泡沫细胞样本筛选得到的风险疾病基因均共享大量的较低层次的分子功能及特异细胞组分(表 1).

Table 1 Common biological function and cellular component of known disease gene and risk disease gene

Gene category	Numbers of gene	Common MF	Common CC
Macrophage	59	55 (93.2%)	24 (40.7%)
Foam cells	61	54 (88.5%)	23 (37.7%)

在动脉粥样硬化病变发展过程中, 血液中的单核细胞与病变区内皮表面黏附分子结合, 在趋化因子作用下, 迁入内膜, 然后转化为巨噬细胞, 进而经特异受体介导的胞吞作用蓄积脂质, 转变成泡沫细胞, 导致斑块表面容易破裂从而引发血栓, 形成动脉粥样硬化的早期病变. 因此我们认为, 巨噬细胞、泡沫细胞对动脉粥样硬化有重要影响.

经过功能富集分析发现, 在巨噬细胞样本、泡沫细胞样本差异表达的基因, 利用支持向量机筛选, 得到的动脉粥样硬化风险疾病基因, 与已知致病基因共享许多重要的生物学功能, 例如: 蛋白质复合体结合、信号转导活性、生长因子活性及细胞因子活性等. 有研究表明, 这些已知致病基因和风险疾病基因共享的生物学功能与动脉粥样硬化密切相关^[12-15]. 例如: 细胞因子在动脉粥样硬化病变过程中的作用是非常复杂的. 巨噬细胞内含有编码组织因子、载脂蛋白 E、转化生长因子以及肿瘤坏死因子的信使 RNA^[16]. 巨噬细胞合成的细胞因子 IL-6 与 TNF 家族等, 可作为炎症及免疫反应的蛋白介质, 以自分泌或旁分泌方式参与白细胞、血管内皮细胞及平滑肌细胞间的信息传递^[17]. 本文筛选得到的巨噬细胞样本、泡沫细胞样本共同风险疾病基因 CSF2RA、CTBP1、IRS2 及 WARS 参与了上述重要细胞过程, 对于动脉粥样硬化的发生发展起着关键作用, 例如其中 CSF2RA 对巨噬细胞有调控作用、控制生长、分化等细胞功能^[18]. 由病变区

巨噬细胞合成的细胞因子介导的细胞间信息传递可能在动脉粥样硬化形成过程中控制着该部位细胞的主要功能。

2.5 信号转导通路分析

利用 Cytoscape 软件中的 CluGO 插件, 对已知疾病基因及风险疾病基因进行信号通路注释分析。发现疾病相关的基因大多分布在低密度脂蛋白通路、NF- κ B 信号通路、生长激素信号通路、外源性凝血酶原激活通路及内脏脂肪沉着综合症等与动脉粥样硬化形成相关的信号转导通路上(图 5)。大量研究表明: CCL2、CSF1、IL6、LDLR 及 LPL 参与的低密度脂蛋白通路, 与动脉粥样硬化病变过程密切相关。巨噬细胞、血管内皮细胞和平滑肌细胞

可把低密度脂蛋白氧化成氧化型低密度脂蛋白, 氧化型低密度脂蛋白可通过巨噬细胞表面的清道夫受体被巨噬细胞摄取, 形成泡沫细胞。氧化型低密度脂蛋白还能吸引单核细胞黏附于血管壁, 对 endothelial 细胞有毒性作用等, 从而促进粥样斑块形成, 进而导致动脉粥样硬化^[19]。研究表明, 胰岛素具有明显的抗炎作用, 可以抑制细胞间黏附分子 ICAM、MCP-1 及 NF- κ B 等炎症因子及黏附分子的表达。本文筛选得到的巨噬细胞样本、泡沫细胞样本共同风险疾病基因 IRS2, 是一类介导胰岛素作用的胞质信号分子^[20], 对于抗动脉粥样硬化形成起到关键作用。

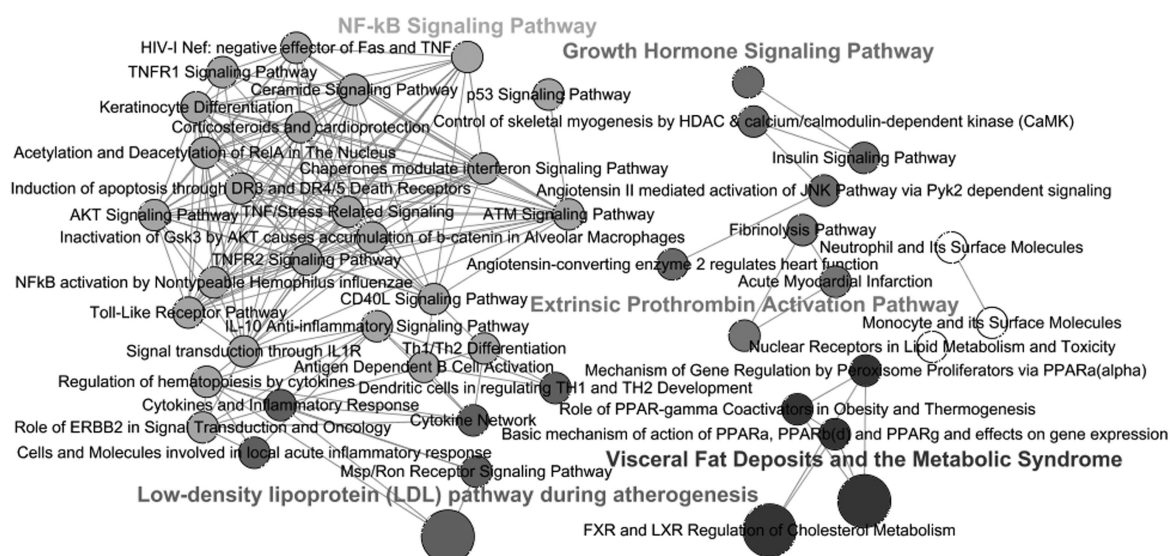


Fig. 5 Signal pathway of disease gene and risk disease gene

3 讨 论

本文提出了基于功能基因组信息、网络拓扑结构信息整合分析的方法, 通过基因组双层筛选, 获得了与动脉粥样硬化病变过程相关的风险疾病基因。此方法弥补了基因芯片表达谱数据分析与蛋白质网络结构单一分析中的不足, 解决了由于高通量实验技术产生较高假阳性率的问题。全面、系统的整合方法, 使我们预测到与动脉粥样硬化发生发展过程密切相关的风险疾病基因。

研究表明, 巨噬细胞、泡沫细胞在动脉粥样硬化病变过程中起到了至关重要的作用。因此对于两类细胞样本, 我们从全基因组层面进行差异表达分

析, 选取在动脉粥样硬化疾病样本与正常样本中显著差异表达的基因作为候选疾病基因(第一次筛选结果)。

通过对蛋白质互作网络中拓扑特征分析, 我们发现, 动脉粥样硬化已知疾病基因趋向于分布在网络中的连接度较高的节点上, 且疾病基因之间是紧密连接的, 通常成簇分布于互作网络中。疾病基因与非疾病基因在度、聚类系数以及最短路径上有较大差异, 因此选择这三个拓扑特征作为支持向量机的分类特征对差异表达基因进行第二次筛选, 最终获得了 59 个巨噬细胞样本中风险疾病基因, 61 个泡沫细胞样本中风险疾病基因。通过基因富集分析及文献验证研究发现, 这些风险疾病基因与已知疾

病基因共同参与了与动脉粥样硬化疾病相关的生物学通路,说明本文的整合双重筛选方法具有较高的准确性。

本方法可以应用到其他复杂疾病致病基因预测、致病通路的挖掘中,为复杂疾病的致病机理研究提供了新的方向。对于复杂疾病的预防、诊断及治疗研究提供了新的理论参考依据。为复杂疾病研究方法由局部向整体,由孤立向系统的转变奠定了坚实的理论与实践基础。

参 考 文 献

- [1] Schadt E E, Lamb J, Yang X, *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 2005, **37**(7): 710–717
- [2] Brent R. Genomic biology. *Cell*, 2000, **100**(1): 169–183
- [3] Sharan R, Suthram S, Kelley R M, *et al.* Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*, 2005, **102**(6): 1974–1979
- [4] Kitano H. Systems biology: a brief overview. *Science*, 2002, **295**(5560):1662–1664
- [5] Peri S, Navarro J D, Kristiansen T Z, *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 2004, **32**(Database issue): D497–501
- [6] Barrett T, Troup D B, Wilhite S E, *et al.* NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, 2007, **35**(Database issue): D760–765
- [7] Edgar R, Domrachev M, Lash A E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 2002, **30**(1): 207–210
- [8] Hamosh A, Scott A F, Amberger J S, *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 2005, **33**(Database issue): D514–517
- [9] Ozgur A, Vu T, Erkan G, *et al.* Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 2008, **24**(13): i277–285
- [10] Ashburner M, Ball C A, Blake J A, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000, **25**(1): 25–29
- [11] BioCarta [DB/OL] <http://www.BioCarta.com>.
- [12] 李 强. 动脉粥样硬化病人增加组织型纤维蛋白溶解酶原激活物抑制物复合体的证据. *国外医学. 老年医学分册*, 1990, **2**(20): 88–89
Li Q. *Foreign Medical Sciences: Gerontology*, 1990, **2**(20): 88–89
- [13] Zhou X, Li D, Yan W, *et al.* Pravastatin prevents aortic atherosclerosis *via* modulation of signal transduction and activation of transcription 3 (STAT3) to attenuate interleukin-6 (IL-6) action in apoE knockout mice. *Int J Mol Sci*, 2008, **9**(11): 2253–2264
- [14] Jackson M, Ahmad Y, Bruce I N, *et al.* Activation of transforming growth factor-beta1 and early atherosclerosis in systemic lupus erythematosus. *Arthritis Res Ther*, 2006, **8**(3): R81
- [15] Loppnow H, Werdan K, Buerke M. Vascular cells contribute to atherosclerosis by cytokine- and innate-immunity-related inflammatory mechanisms. *Innate Immun*, 2008, **14**(2): 63–87
- [16] Wilcox J N, Nelken N A, Coughlin S R, *et al.* Local expression of inflammatory cytokines in human atherosclerotic plaques. *J Atheroscler Thromb*, 1994, **1**(Suppl 1): S10–13
- [17] 秦树存, 王士雯. 细胞因子与动脉粥样硬化. *国外医学. 老年医学分册*, 1994, **15**(5): 193–196
Qin S C, Wang S W. *Foreign Medical Sciences: Gerontology*, 1994, **15**(5): 193–196
- [18] Trapnell B C, Carey B C, Uchida K, *et al.* Pulmonary alveolar proteinosis, a primary immunodeficiency of impaired GM-CSF stimulation of macrophages. *Curr Opin Immunol*, 2009, **21**(5): 514–521
- [19] 胡桂才, 唐世英, 李春华, 等. 泡沫细胞与动脉粥样硬化. *承德医学院学报*, 2008, **25**(1): 81–83
Hu G C, Tang S Y, Li C H, *et al.* *J Chengde Medical College*, 2008, **25**(1): 81–83
- [20] Razani B, Chakravarthy M V, Semenkovich C F. Insulin resistance and atherosclerosis. *Endocrinol Metab Clin North Am*, 2008, **37**(3): 603–621

Uncovering Atherosclerotic Risk Disease Gene Based on Expression and Network Topological Structure*

WANG Hong, QU Xiao-Li, ZHAO Yan, ZHANG Jing, CHEN Li-Na**

(College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China)

Abstract The integrated analysis method based on functional genome information and network topological structure information mining atherosclerotic risk disease gene, provide a new perspective of atherosclerosis research in genome level. Through double selection, differential expression analysis and support vector machine (SVM), disease related risk gene with high confidence could be uncovered. It is helpful for accurately distinguishing the disease genes and non-disease genes in the protein-protein interaction network, and building relationship between two type of these genes. Base on our strategy, 59 risk disease genes were exploited from macrophages sample and 61 risk disease genes from foam cell sample which shared common biological function and signal pathways with known AS disease gene. Furthermore, this method could be used to study the pathogenesis of other complex diseases in genome level.

Key words atherosclerosis, differential expression analysis, network topological characteristic, support vector machine

DOI: 10.3724/SP.J.1206.2009.00687

*This work was supported by grants from The Master Innovation Funds of Harbin Medical University (HCXS2010006), The National Science Foundation of Heilongjiang Province (D2007-48), The National Natural Science Foundation of China (30571034) and The Undergraduate Innovation Funds of Harbin Medical University (HRBCX2008-02).

**Corresponding author.

Tel: 86-451-86674768, E-mail: chenlina_2004@yahoo.com.cn

Received: March 25, 2010 Accepted: May 17, 2010