

基于功能一致性和网络拓扑属性 预测冠心病致病基因*

赵 研 陈丽娜** 张良才 王 倩 尚玉奎 王 宏 李 琬

(哈尔滨医科大学生物信息科学与技术学院, 哈尔滨 150081)

摘要 基于功能一致性利用蛋白质互作网络挖掘潜在的疾病致病基因, 对于了解疾病致病机理和改进临床治疗至关重要. 基于基因功能一致性和其在蛋白质互作网络中的拓扑属性将基因与疾病之间建立关联, 对疾病风险位点内的基因进行了致病风险预测, 并通过 GO 及 KEGG 功能富集分析方法进一步筛选, 预测出新的致病基因. 预测出了 51 个新的冠心病致病基因, 分析发现大部分基因参与了冠心病的致病过程. 为疾病基因的挖掘提出一个新的思路, 从而有助于复杂疾病致病机理的研究.

关键词 网络拓扑性质, 共定位特性, 功能富集分析, 功能一致性

学科分类号 R318.04, Q78

DOI: 10.3724/SP.J.1206.2008.00623

复杂疾病被认为是由许多个基因相互作用所导致的疾病. 人们研究复杂疾病的一个主要目标是寻找与疾病相关的致病基因并且阐释它们在疾病的发生发展中的作用. 推测致病基因是一个在复杂疾病研究中的重要挑战, 许多的方法被用来挖掘复杂疾病相关的候选基因. Schwikowski 等^[1]和 Titz 等^[2]提出蛋白质同它们直接相互作用的邻居之间具有相似的功能. Jonsson 和 Bates^[3]研究发现, 致病基因之间的互作要高于非致病基因之间的互作 2 倍或更多, 致病基因比非致病基因在网络中更倾向于成簇存在. 在蛋白质互作网络中, 基于网络拓扑属性研究疾病与基因之间关系的方法被广泛地应用到疾病致病基因的预测研究中. 但是, 单纯从网络拓扑性质不能全面准确地反映基因与疾病之间的真实关系. 蛋白质的功能一致性不仅体现在网络拓扑性质的一致上, 而且体现在编码这些蛋白质的基因在序列上所表现出的共定位特性^[4]. Oti 等^[5]和 Chen 等^[6]利用在蛋白质互作网中与已知的疾病基因直接互作的关系, 对 OMIM 数据库中的疾病显著风险位点中的致病基因进行了筛选并成功预测出了新的遗传异质性疾病致病基因.

现今, 许多生物学数据库提供了基因所具有的

生物学功能信息. 例如, Gene Ontology(GO)^[7,8]数据库包含了基因参与的生物过程, 所处的细胞位置及具有的分子功能三方面功能信息, 通过 GO 中的注释信息, 我们可以对基因的功能进行功能分析. 京都基因和基因组百科全书(KEGG)^[9]是系统分析基因功能、联系基因组信息和功能信息的知识库, KEGG 的 PATHWAY 数据库提供了基因编码的生物学大分子酶或者蛋白质在生命体内相互联系相互影响的情况. 同一生物学通路内的基因大多参与了此代谢通路所揭示的生命过程.

根据功能相似的基因可能导致相似的疾病表型^[10,11]这一理论依据, 通过网络拓扑性质对疾病风险位点上的基因与疾病之间的关联关系进行预测, 并利用 GO 和 KEGG 功能富集分析方法进一步筛选预测出新的致病基因.

* 哈尔滨医科大学研究生创新基金(HCXS 2008010), 黑龙江省自然科学基金(D2007-48), 哈尔滨医科大学医学基础学科创新群体基金和国家自然科学基金(30571034)资助项目.

** 通讯联系人.

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

收稿日期: 2008-09-08, 接受日期: 2008-10-30

1 材料与方法

1.1 数据源

通过 OMIM 数据库, 共获得已知的冠心病相关的致病基因共 138 个及相关风险条带共 168 条. 全基因组基因在染色体上的定位信息和全基因组染色体条带定位信息来自于 Ensembl 数据库. 人类蛋白质互作数据来源于 HPRD 数据库 (下载时间为 2007 年 9 月). 对下载的 HPRD 互作数据进行整理

后, 得到一个包括 9 013 个蛋白质, 34 032 个非自身互作关系的人类蛋白质 - 蛋白质互作网络. 相应的互作数据下载网址为: <http://www.hprd.org/download>. GO 数据库信息和 KEGG 通路信息下载地址为 <http://geneontology.org/> 和 <http://www.genome.jp/kegg/>.

基于功能一致性及网络拓扑属性挖掘疾病致病基因具体流程如下(图 1):

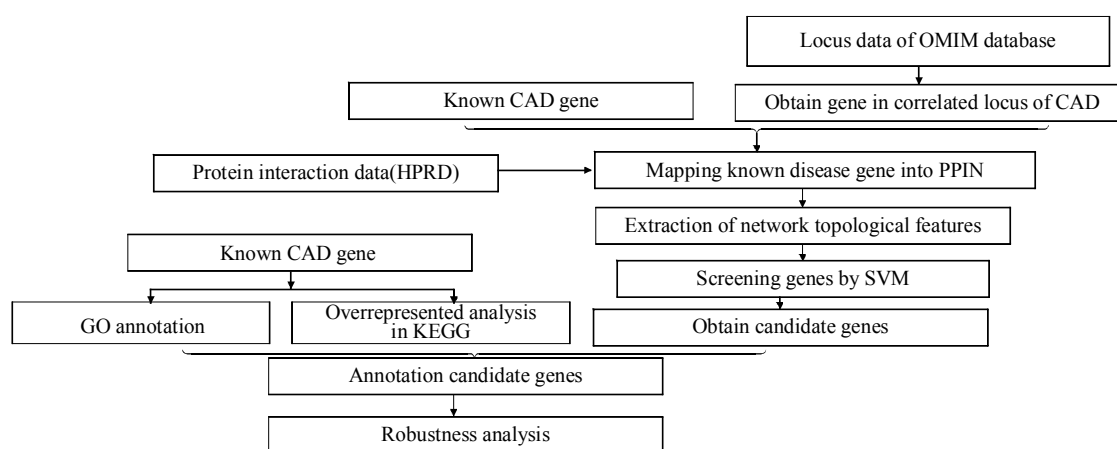


Fig. 1 The dataflow of our method

1.2 获取已知的冠心病风险条带上的基因

通过文本挖掘方法, 我们从 OMIM 数据库中获取了 168 条与冠心病相关的风险条带的编号、染色体位置、起始位点和终止位点位置等信息. 随后, 我们利用序列共定位方法对蛋白质互作网络中的基因进行了筛选. 如果一个基因在染色体中的位置处于已知的风险条带的起止位置附近上下游 500 kb 范围内, 就认为这个基因是一个候选基因, 构成候选基因集合 A.

1.3 利用基因在网络中的拓扑属性获取疾病候选基因集合

从 OMIM 数据库中我们获得了与冠心病相关的已知疾病基因共 138 个, 构成了已知致病基因集 M. 本文提出 4 种网络拓扑属性来描述基因在网络中的拓扑特征, 分别为基因在网络中的度、已知致病基因邻居的个数、聚类系数及与已知致病基因间最短路径的均数. 利用支持向量机对候选基因集合 A 中的每一个基因进行疾病关联性预测. 我们只保留候选基因集合 A 中在预测过程中被分类为疾病基因的候选基因作为疾病风险基因, 从而对候选基

因集合 A 进行筛选, 最后得到疾病风险基因集合 B.

1.4 疾病风险基因集合 B 的功能筛选和新的致病基因集的确定

利用 Gene Ontology 和 KEGG 数据库, 对冠心病已知致病基因集 M 和疾病候选基因集 B 分别进行生物学功能、细胞过程和细胞组分、生物学通路功能注释. 根据疾病致病基因之间往往具有相同 GO 功能和共享相同通路, 筛选候选基因集 B 中与已知致病基因集 M 功能一致的基因, 作为冠心病新的致病基因集.

1.5 疾病基因的网络特征

以人类蛋白质互作网络为背景, 将预测的冠心病新的致病基因和已知致病基因均作为冠心病疾病基因, 分析其在蛋白质互作网络中的拓扑特征, 寻找冠心病疾病基因在网络中的一般属性.

2 结 果

2.1 利用网络拓扑性质确定基因与疾病间的关联关系

通过对 OMIM 数据库的检索, 我们得到了 168

条与冠心病相关的风险条带，利用冠心病风险条带信息对蛋白质互作网络中的基因进行筛选，只保留位于风险条带附近 500 kb 范围内的基因作为疾病候选基因。随后，通过分析已知疾病基因在网络中的拓扑特征，利用支持向量机对疾病候补基因进行分类，结果发现在 4 种网络拓扑属性中，利用已知疾病基因邻居的个数来预测疾病基因取得了最好的预测效果。最终选取分类得到的疾病基因作为初步预测的冠心病候补基因，共得到 247 个基因。

2.2 已知致病基因与候选基因功能注释分析

为了对得到的 247 个预测的冠心病候补基因进行筛选，我们以已知致病基因所富集的 GO 功能类及其所在的 KEGG 通路为背景，利用超几何分布富集分析(显著性水平为 0.01)对预测的疾病候选基

因进行进一步富集分析。通过对已知致病基因的 GO 功能富集分析发现：已知致病基因大多分布在发育、生长和应激反应三个生物学过程；这些基因所编码的蛋白质都倾向于在胞外区域行使抗氧化剂活性、结合绑定和信号转导活性三大生物学功能(图 2)。随后我们发现，预测的 247 个候选致病基因中有 98 个基因注释到了已知的致病基因所富集的 GO 注释上。同时通过对 247 个候选致病基因在已知致病基因所富集的 KEGG 通路上进行富集后发现，其中有 111 个基因共享了已知致病基因所富集的黏着斑通路、肌动蛋白细胞骨架调节通路、细胞因子受体互作通路和 MAPK 代谢通路等生物学通路(图 3)。

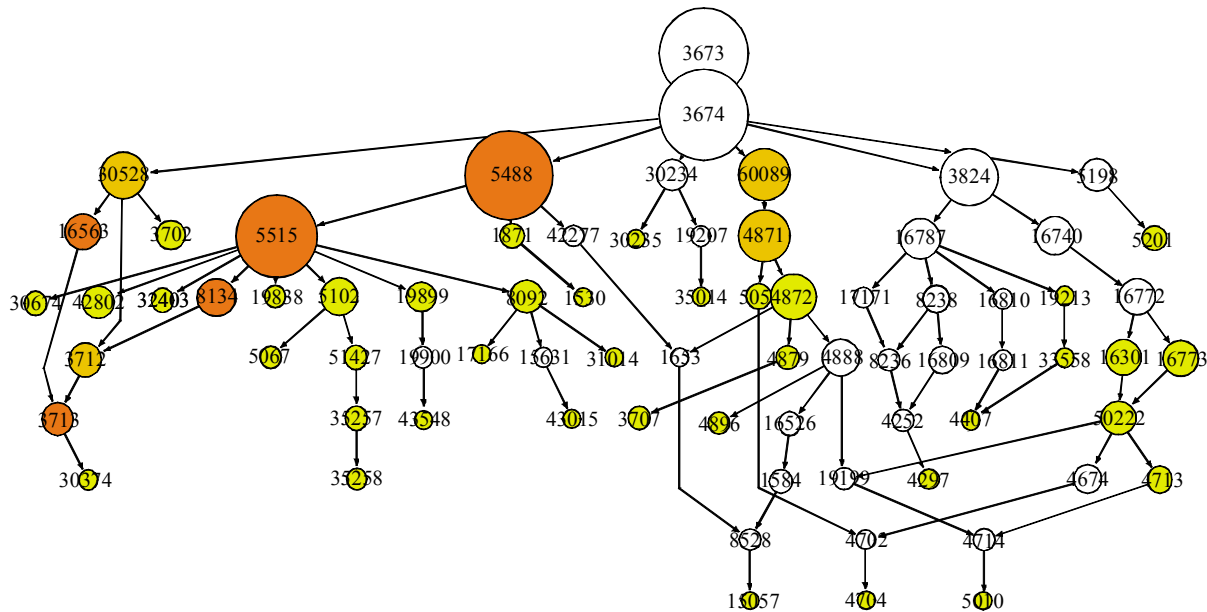


Fig. 2 The GO function overrepresent analysis of specific disease genes of coronary artery disease
The nodes in yellow are overrepresented by specific disease-genes of coronary artery disease in the figure. The more dark nodes represent the GO terms are overrepresent by more specific disease-genes of coronary artery disease.

因此，为了保证预测结果的准确性，我们保留与已知致病基因共享 GO 注释且具有相同的 KEGG 生物学通路的基因作为预测发现的冠心病致病基因，最终挖掘得到了 63 个冠心病致病基因，其中包括 12 个已知致病基因和 51 个预测的候补致病基因。

2.3 冠心病致病基因的网络拓扑性质

以人类蛋白质互作网络为背景，对 51 个新的冠心病致病基因及已知的冠心病致病基因在网络中的拓扑特征进行了分析(图 4)。从图 4 中可以看出，同网络中的其他基因相比，冠心病致病基因在网络中倾向于分布在网络中高连接的节点上，冠心病致病基因在网络中一般具有较多的互作关系。

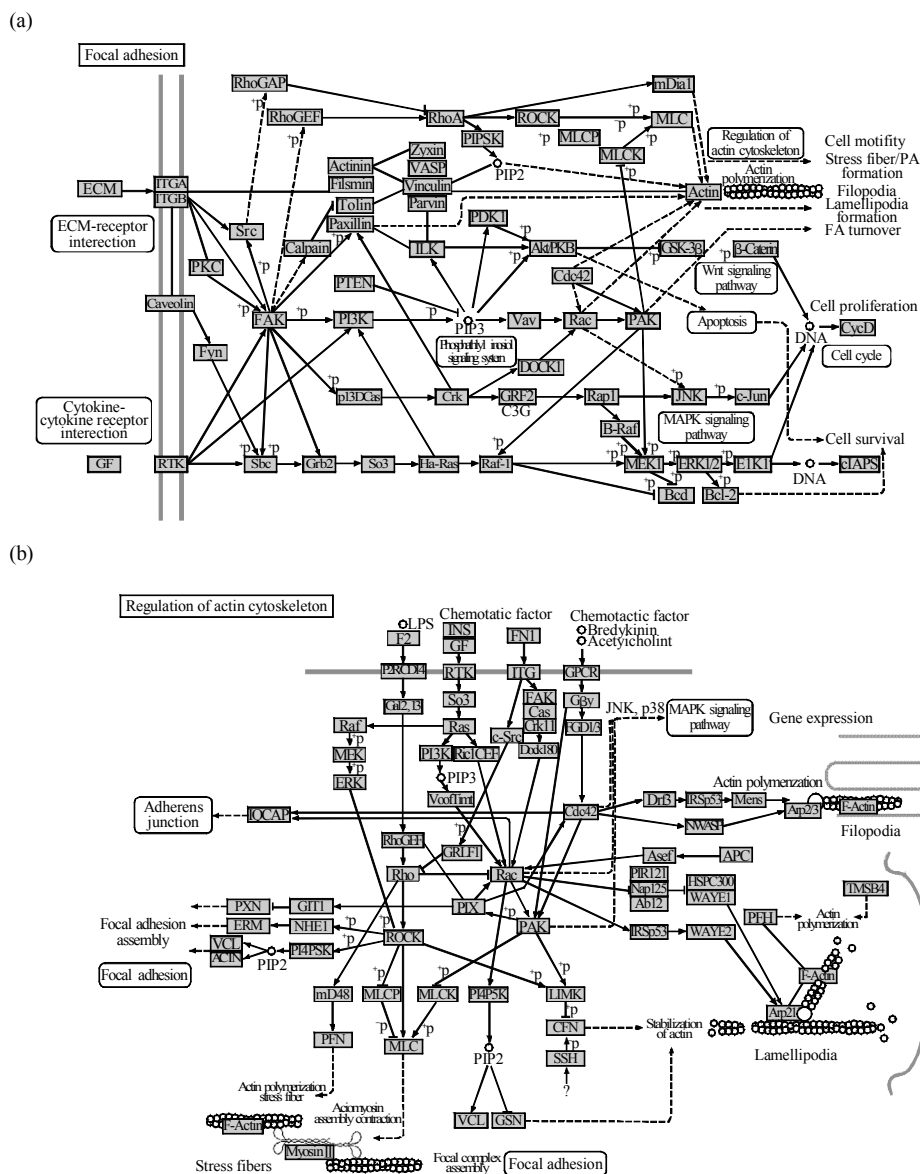


Fig. 3 The KEGG pathway enriched by known and candidate disease genes

(a) Pathway of focal adhesion. (b) Pathway of regulation of cytoskeleton.

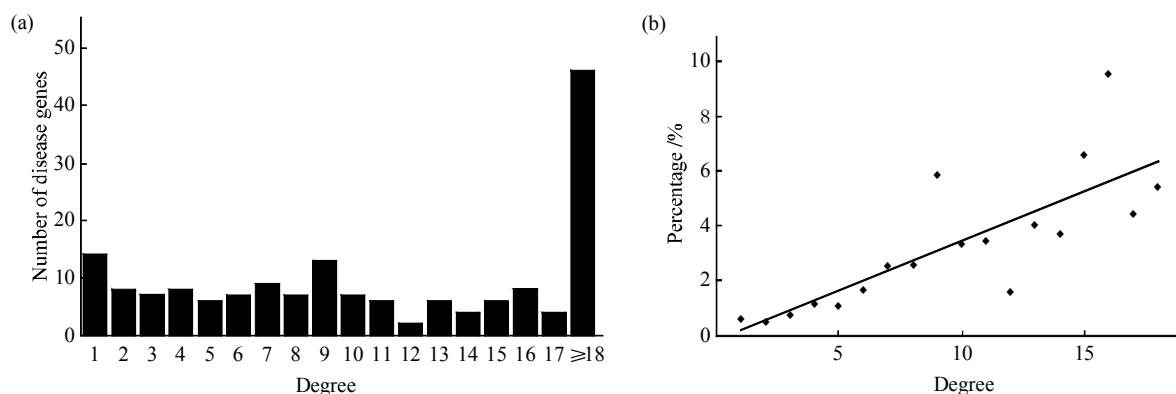


Fig. 4 Analysis of the network topological features of disease genes for coronary artery disease

(a) Degree-number of disease genes. (b) Degree-the percentage of disease genes in all network genes.

3 讨 论

疾病基因预测是复杂疾病致病机理研究工作中的一个重要内容。基于蛋白质网络拓扑特征及生物学功能一致性来预测复杂疾病的候选致病基因。选择已知的疾病基因邻居的个数作为筛选特征, 利用支持向量机对疾病风险位点上的基因进行了疾病关联性预测, 结果发现有 249 个疾病候选基因。随后, 为了保证预测的疾病基因的准确性, 基于功能相似的基因通常导致表型的一致这一理论, 我们对预测得到的疾病候选基因和已知致病基因进行 GO 功能注释和 KEGG 通路富集分析。根据功能一致性, 按照预测的基因和已知致病基因共享相同的 GO 功能类和共出现的 KEGG 通路这一标准, 对预测的候选基因进行筛选, 并最终得到了 51 个新的冠心病致病基因。本文研究发现, 同蛋白质互作网络中的非致病基因相比, 心血管疾病基因在网络中多分布在连接度较高的节点上, 是蛋白质互作网络中重要的组成部分, 它们在实现蛋白质功能和生命活动中发挥着关键作用。这为研究人员挖掘冠心病疾病基因并筛选冠心病药物靶点提供了参考。而冠心病疾病基因的高度互作同时表明冠心病的发生不是由单个基因或一小群基因所影响的, 是由于大量的疾病相关基因共同参与, 相互作用所导致的。

随后通过文献挖掘, 我们对预测得到的 51 个冠心病致病基因进行了验证, 结果发现大部分基因都与冠心病的发生发展有着密切的联系: He 等^[12]认为冠心病的发生发展与异常的凝血系统有着紧密的关系, 利用酶关联的免疫吸收实验对 54 个冠心病病人血浆内蛋白 C、游离蛋白 S、激活因子Ⅶ和纤溶酶原活化因子等生物分子水平进行鉴定, 研究发现, 血浆中蛋白 C 的活性降低、凝血因子和激活因子Ⅶ的活性增强都很容易促成冠心病的发生发展。Elbatarny 等^[13]研究表明 LEPR 基因编码的 Leptin 受体是心血管疾病的一个独立的风险因子。研究发现, leptin 通过激活信号级联反应能够诱导血小板活性, 促进 ADP 诱导的血小板聚集, 从而导致血栓形成, 引起急性冠心病的发作。血小板衍生生长因子-BB 和胰岛素生长因子-1 以及它们的受体之间的相互作用调控了平滑肌的增殖, 从而引

起冠状动脉粥样硬化。

因此, 本文提出了一种新的疾病基因的预测方法, 并将其成功地应用到冠心病疾病基因的预测研究中并得到了很好的研究结果。本文的研究不仅仅为复杂疾病致病基因的预测提供了一个新的方法, 而且通过对疾病基因的分析发现了冠心病疾病基因的网络属性, 为冠心病的致病机理的研究提供了指导意义。

参 考 文 献

- Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 2000, **18**(12): 1257~1261
- Titz B, Schlesner M, Uetz P. What do we learn from high-throughput protein interaction data?. *Expert Rev Proteomics*, 2004, **1**(1): 111~121
- Jonsson P F, Bates P A. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 2006, **22**(18): 2291~2297
- Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 2007, **8**(12): 995~1005
- Oti M, Snel B, Huynen M A, *et al.* Predicting disease genes using protein-protein interactions. *J Med Genet*, 2006, **43**(8): 691~698
- Chen J Y, Shen C, Sivachenko A Y. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput*, 2006, **11**: 367~378
- The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet*, 2000, **25**: 25~29
- Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 2005, **21**(16): 3448~3449
- Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl Acid Res*, 2000, **28**: 27~30
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*, 1996, **273**(5281): 1516~1517
- 孙玉琳, 赵晓航. 复杂疾病基因定位策略与肿瘤易感基因鉴定. *生物化学与生物物理进展*, 2005, **32**(9): 803~810
- Sun Y L, Zhao X H. *Prog Biochem Biophys*, 2005, **32**(9): 803~810
- He W J, Hu Y, Zhang X P, *et al.* Activated protein C ratio, plasma tissue factor activity and activated factorⅦ in Chinese patients with coronary heart disease. *Eur J Med Res*, 2008, **13**(2): 47~51
- Elbatarny H S, Maurice D H. Leptin-mediated activation of human platelets: involvement of a leptin receptor and phosphodiesterase 3A-containing cellular signaling complex. *Am J Physiol Endocrinol Metab*, 2005, **289**(4): E695~702

Predicting Disease Genes of Coronary Artery Disease Based on Functional Consistency and Network Topological Features*

ZHAO Yan, CHEN Li-Na**, ZHANG Liang-Cai, WANG Qian, SHANG Yu-Kui, WANG Hong, LI Wan

(College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China)

Abstract The identification of genes responsible for human diseases based on functional consistency and network topological features is of great importance for both understanding human disease pathogenesis and improving clinical practice. A novel method based on the functional consistency and network topological features was introduced to establish an association between genes and diseases. Using this method, candidate disease genes were predicted from disease risk loci. Then, the candidate genes sharing the same or similar functions with known disease genes in the functional enrichment analysis of GeneOntology and KEGG databases as final disease genes were determined. 51 genes were predicted to be the disease genes for coronary artery disease and most of them participate in the development of disease by literature retrieval. The method provided additional insights for the finding disease genes, which will be helpful for the studies on the pathogenesis of human complex diseases.

Key words network topological features, co-located attribution, functional enrichment, functional consistency
DOI: 10.3724/SP.J.1206.2008.00623

*This work was supported in part by grants from The Master Innovation Funds of Harbin Medical University (HCXS 2008010), The National Science Foundation of Heilongjiang Province (D2007-48), The National Natural Science Foundation of China (30571034) and The Groups Innovation Funds of Harbin Medical University.

**Corresponding author.

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

Received: September 8, 2008 Accepted: October 30, 2008