doi: 10.13241/j.cnki.pmb.2014.13.041

·生物信息学·

融合基因关系网和功能预测恶性胶质瘤基因方法研究 *

孙红梅[1] 常志强[1] 张淑娟[2] 许 艳[1△]

（1 哈尔滨医科大学生物信息科学与技术学院 黑龙江 哈尔滨 150086；
2 哈尔滨医科大学基础医学院 黑龙江 哈尔滨 150086）

摘要 目的：建立挖掘恶性胶质瘤候选基因的方法并进行系统分析。**方法：**结合恶性胶质瘤已知通路内基因和发生点突变和拷贝数改变的基因构建扩展基因关系网络，计算并分别寻找在网络中度和中心性得分高，脆弱性为正数的节点（基因），将满足一种或多种测度并与已知恶性胶质瘤基因共功能的基因作为恶性胶质瘤候选基因。最后，通过文献验证方法评价多种测度预测恶性胶质瘤基因的效能。**结果：**融合基因功能后，利用基因在网络中的度和脆弱性可识别大部分恶性胶质瘤基因，但利用中心性预测的结果较差；当将三个测度融合后，效能并没比单独使用脆弱性高。**结论：**融合基因功能关系和网络脆弱性是预测恶性胶质瘤基因的最佳测度。

关键词：基因关系网络；脆弱性；中心性；恶性胶质瘤

中图分类号：R739.41 **文献标识码：**B **文章编号：**1673-6273(2014)13-2549-05

Research of Prediction of GBM Genes by Integrating Gene Relational Network and Gene Function*

SUN Hong-mei[1], CHANG Zhi-qiang[1], ZHANG Shu-juan[2], XU Yan[1△]

*(1 College of Bioinformatics Science and Technology, Harbin Medical University, Heilongjiang, Harbin, 150081, China;
2 Basic Medical Science College, Harbin Medical University, Heilongjiang, Harbin, 150081, China)*

ABSTRACT Objective: To establish and analysis the method of mining malignant glioma candidate genes. **Methods:** Combining the pathway of malignant glioma, point mutations and copy number changes in genes, based on the gene function and degree, vulnerability and the centrality of genes in network, this study analysis and evaluate the effectiveness of these characteristics in the identification of malignant glial the effectiveness of tumor related genes. **Results:** Combining the gene function, the degree and vulnerability of the genes in the network can identify most of the malignant glioma gene, while central forecast is poor. When combining three measures, the performance did not perform better than the use of vulnerability. **Conclusion:** The gene functional relationships and network vulnerability is the best measure to predict malignant glioma genes.

Key words: Gene relational network; Vulnerability; Centrality; Glioblastoma

Chinese Library Classification: R739.41 Document code: B

Article ID: 1673-6273(2014)13-2549-05

前言

恶性胶质瘤（glioblastoma，GBM）是最常见的原发性颅腔内恶性肿瘤[1,2]。根据它的病理情况可分为少枝胶母细胞瘤、多形胶母细胞瘤、星形细胞瘤、髓母细胞瘤、室管膜瘤等。近年来，恶性胶质瘤的发生率在逐年增加，年增长率大概为 1.2%。目前，一般用手术切除和放疗辅助等方法对恶性胶质瘤的患者进行治疗，尽管如此，患者的平均寿命大约为 1 年。因此，目前仍然需要寻找新的恶性胶质瘤治疗方案。研究显示，恶性胶质瘤的发病机制及特异分子网络主要涉及了 PI3K-AKT-mTOR 和 Ras-MAPK 信号通路与受体酪氨酸激酶信号通路之间的作用，以及信号通路 sonic hedgehog (SHH)和 Wnt 与调节细胞生长周期的 p53 和 RB 信号通路之间的作用 [3]。与此类似，ARF 和 INK4 作为 p53 和 RB 的重要关键调节因子，由于编码其 CDKN2A 位点发生缺失或突变，p53 拮抗剂 MDM4 和 MDM2 也发生了片段的扩增[4]。在最近研究中，从基因 CDKN2A 和临近基因 CDKN2B 的位点中识别出与胶质瘤发展有关的风险单核苷酸多肽 [5]，还识别出与其关联的基因，如端粒酶逆转录酶(TERT)和端粒延长螺旋酶 1 的调节因子（RTEL1）[6]。

研究发现，恶性胶质瘤有多个靶基因，这些基因都依赖细胞而长期存活[7-9]。我国的刘福生等人[10]将病毒基因组的 ICP6 和 ICP34.5 基因敲除，插入到人的 Endo Angio 融合基因进行改

* 基金项目：黑龙江省自然科学基金资助项目（D201116）

作者简介：孙红梅（1982-），女，硕士研究生，讲师，研究方向：肿瘤生物信息学，分子遗传学，电话：0451-86669617，
E-mail: sunhongmei@hrbmu.edu.cn
△通讯作者：许艳，E-mail: xuyanls@yahoo.com.cn

（收稿日期：2013-11-21 接受日期：2013-12-19）

造，使该病毒拥有溶解肿瘤细胞的溶瘤性质，且肿瘤细胞能够在溶解前表达 Endo Angio 基因，从而抑制了血管的生成，达到治疗效果。此外，基因 Ras 在胶质瘤的恶性侵袭、转化生长中起着关键的作用[11]。这些研究结果表明，在恶性胶质瘤中，破坏基因或信号通路会对肿瘤细胞的周期调节有重大影响。寻找靶基因是治疗疾病和研发药物的重要步骤。目前，许多研究者致力于寻找疾病的靶基因[12]。恶性胶质瘤对人类的危害促使人们要更全面的了解其致病机理，识别候选致病基因。生物网络能够表示基因在疾病特异网络中的属性，亦能描述出基因在网络中的作用。此外，疾病相关的基因倾向于具有共同的生物学功能。本文将结合基因在基因关系网络中的属性以及恶性胶质瘤基因的功能来预测疾病基因。为了评价预测方法的有效性，本文分别计算了网络中节点的度、脆弱性和中心性三种测度与基因功能结合后再识别新基因与已知基因的比例，经过比较分析，筛选出最优的预测方法。该方法不仅适用于识别已知的疾病基因，也可以根据已知基因预测新的疾病基因，为未来的疾病基因预测提供良好的模板和思路。

1 材料与方法

1.1 材料

利用文献和公共通路数据库，癌症基因组学（CancerGenomics）整理了一个在神经胶质瘤中最频繁改变的基因构成的通路，共包含了 73 个基因。神经胶质瘤的 DNA 拷贝数变异数据来自基因组内显著的扩增和缺失片段[13]。所有这些数据均从 CTGA 数据中下载，网址为 http://cbio.mskcc.org/cancergenomics/gbm/。

基因互作关系对来自当前通路数据库和医学文献，包括 KEGG，BioCarta，NCI-Curated。文献中的基因 / 蛋白质之间的关系来自文本挖掘算法从医学文献数据库中挖掘得到的基因间的关系，主要的关系类型包含激活、抑制、调控等。

1.2 方法

1.2.1 **基因网络中筛选 GBM 候选基因** 为了扩增 GBM 相关的候选基因，本文从生物学网络中选取具有特定属性的基因。涉及的属性包括基因的度(Degree, D)，基因的脆弱性(Vulnerability, V)和基因的中心性(Closeness centrality, CC)。基因的脆弱性是用网络的全网效率计算得到的。全网效率衡量了网络在传播节点间信息的有效性，其中任意两个节点间的有效性与他们的距离成反比[14]。网络中第个节点的脆弱性与网络没有第个节点的全网效率有关。如果越大，表示删除第个节点对网络的全网效率的影响越大。

$$E=\frac{1}{N(N-1)}\sum_{i\neq j}\frac{1}{d_{ij}} \qquad V_i=(E-E_i)/E \qquad (1)$$

其中为第个和第个节点间最短路径中边的数值，为网络中节点的个数。

中心性描述了某个节点与其它节点靠近的紧密程度，所以它可以作为基因在网络中是否处于中心位置的衡量指标。第个节点的中心性用下面公式进行计算。

$$Closenesscentrality_i=\frac{|V_i|-1}{\sum_{i\neq j}d_{ij}} \qquad (2)$$

为了从生物网络中选取 GBM 相关的候选基因，在保证网络中节点度不变的情况下，随机扰动网络中的边，计算每个基因在网络中的脆弱性和中心性。经过 1000 次随机后，运用下列原则选取 GBM 相关的候选基因：1.在 GBM 基因网络中，至少与两个 GBM 相关基因直接相连的基因，或者与某一 GBM 相关基因的关系出现在两个以上数据集合中的基因；2.对全网效率有正贡献的基因（脆弱性得分 >0）；3.接近网络中心性得分排名在前 50 位的基因。4.选取在网络中度排名前 50 的基因。

1.2.2 **筛选 GBM 基因共功能的基因** 为了找出与 GBM 相关基因具有相同或类似功能的基因，本文将得到的 GBM 相关基因映射到 GO 数据库中的生物学过程、细胞组分、和分子功能中，用超几何分布计算 GBM 基因显著富集在 GO 条目中的显著性 p 值，通过运用 Benjamini-Hochberg 校正后，选取错误发现率（FDR）小于 0.01 的 GO 条目中的基因作为 GBM 相关的候选基因。

$$P=1-\sum_{i=0}^{x}\frac{\binom{k}{i}\binom{M-K}{N-i}}{\binom{M}{N}} \qquad (3)$$

1.2.3 GBM **相关基因预测** 已有研究表明，与某疾病有关的基因在其生物网络中有更高的度[15,16]；在生物网络中，具有较高脆弱性和中心性的基因更倾向于与疾病相关[17,18]；此外，从基因功能角度讲，与疾病基因有相同的功能的基因更可能与疾病有关[19,20]。基于这几个已经证实的假设，本文将来自网络中的三个属性的基因集合合并，然后将结果与功能注释得到的基因取交集，可以认为通过这两个测度筛选出的基因更加倾向于与 GBM 相关。

1.2.4 **预测方法检验** 为了验证预测 GBM 相关基因的方法是否有效，本文从 2 个角度进行验证。首先计算预测的 GBM 相关基因与真实基因重复比率，衡量方法在挖掘已知疾病基因方面的效能。其次，通过查找已有的医学文献，验证预测出的非已知 GBM 基因是否与 GBM 的有关并计算比例。

2 结果

2.1 GBM 相关数据

本文从 CTGA 中共搜集了 601 个基因，其中来自通路的基因有 73 个，具有 2 个或更多错义突变的基因 42 个，来自 DNA 拷贝数变异的基因 486 个。通过从 KEGG、BioCarta、NCI Nature Curated 和 MRDB 数据库中整理基因间的关系，提取至少包含一个 GBM 相关基因关系构建 GBM 相关基因关系网络，同时保证其中的非 GBM 基因至少与 2 个 GBM 相关基因有连接。

2.2 GBM 基因网络属性

GBM 基因网络共包含了 1109 个基因，4131 对基因关系。该网络的聚类系数为 0.084，比随机情况下的聚类系数高 0.03± 0.005（P<0.01），这说明 GBM 基因网络的聚集性要明显优于正常情况下的网络聚集性；网络的度分布服从 PowerLaw 分布，说明 GBM 基因网络符合生物学网络的特性。通过计算网络中每个节点的脆弱性和中心性，比较排除每个节点后全网效率和排除前的全网效率，共得到了 54 个全网效率有正效应

的基因。加上度按大小排序前 50 的基因,中心性排列前 50 的基因，共计 154 个基因。脆弱性为正且度排列前 50 的基因共 73 个,31 个重复；脆弱性得分为正值且结合中心性排前 50 的基因共 104 个,0 个重复；度和中心性排前 50 的基因有 100 个,0 个重复。

2.3 具有 GBM 相关功能的候选基因筛选

通过将 GBM 相关基因映射到 GO 数据库中,用超几何检验计算 GBM 基因与每个 GO 条目内基因重复的显著性,选取 BH 校正后 FDR 小于 0.01 的 GO 功能条目作为 GBM 的相关条目,共得到了 41 个显著的 GO 条目,合计 4058 个基因。这 41 个 GO 条目中有 7 个与细胞自身活性有关,10 个与磷酸化作用有关。

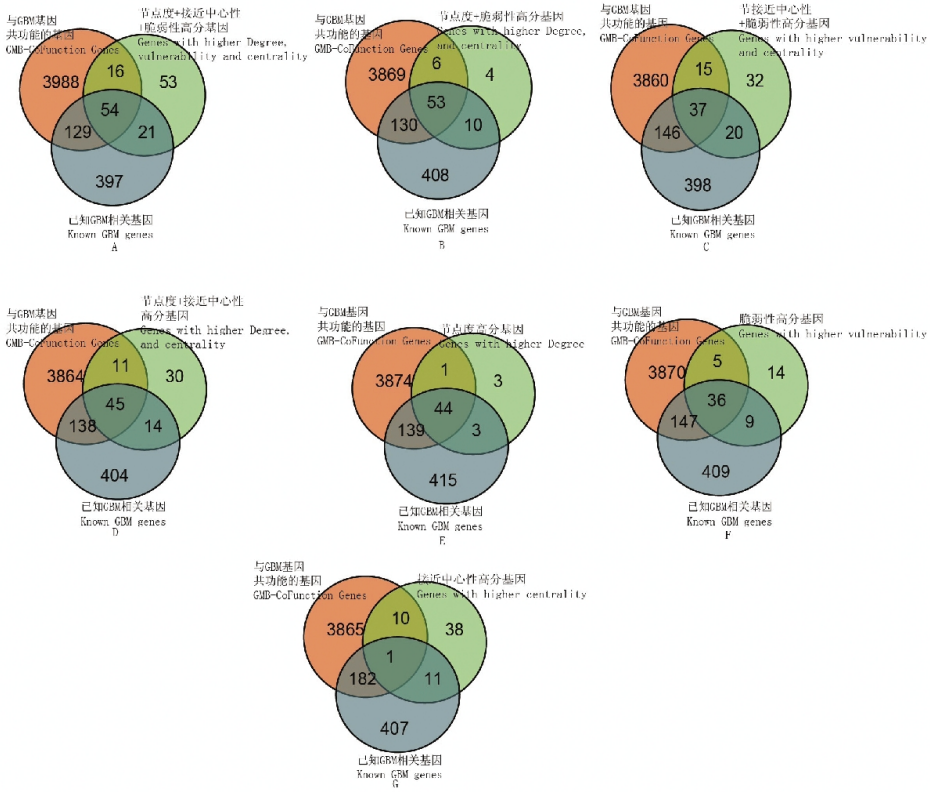

图 1 不同方法预测 GBM 基因结果

Fig. 1 The result of predicting GBM genes with different methods

2.4 GBM 候选基因预测

为了寻找用于基因预测的最佳网络属性,本文比较了多种测度预测 GBM 基因的正确率。结果见图 1,与三种测度相关且与 GBM 共功能的基因有 70 个，其中 54 个基因是已知的 GBM 基因(表 1)。此外,本文计算了结果中含已知 GBM 基因的比例,如果该比例越高,说明预测已知 GBM 基因的概率越大,其预测出新 GBM 基因的效能也就越小。

节点的度测度结合共功能共预测出 44 个 GBM 已知基因(预测率(P_original):97.78%),而节点的脆弱性、度测度结合共功能预测出 36 个 GBM 已知基因 (89.83%)(表 1)。结果表明,在已知基因构建的 GBM 基因网络中,GBM 基因更加倾向于具有较高的度、脆弱性。相反,利用节点的中心性作为测度预测出 GBM 基因只有 1 个(预测率为 9.09%),表明节点的中心性并不能较好的作为预测疾病基因的测度。从另一个角度看,如果某个测度对已知 GBM 基因的预测率越高,则其对候选的 GBM 基因的预测率越低。节点的度结合共功能仅仅预测出 1 个候选 GBM 基因,用脆弱性测度预测出 5 个候选 GBM 基因,而度结合脆弱性测度预测出 6 个候选 GBM 基因,这些都比其它测度预测出候选 GBM 基因数目少。

表 1 不同测度预测 GBM 基因的结果

Table 1 The result of different measures on predicting GBM genes

	DCV	DV	CV	DC	V	D	CC
Predicted Gene	70	59	52	56	41	45	11
GBM Genes Number	54	53	37	45	36	44	10
GBM Gene Ratio(%)	77.14	89.83	71.15	80.35	87.80	97.78	9.09

Note: DC:D+CC; DV:D+V; DCV:D+CC+V; CV: C+V

2.5 网络测度预测新基因能力

为了验证三种测度分别结合功能预测的基因是否与 GBM 有关，本文将每个基因的名字、别名和神经胶质瘤作为关键字，检索医学文献，选取包含这些关键字的文章进行人工验证，考查他们是否在同一个句子中出现，或者有关联。在综合度、中心性和脆弱性测度时，本文预测出的基因中有 16 个得到文献支持（表 2），其中 8 个基因与 GBM 有直接的关联；只综合度和脆弱性两个测度预测出的 6 个基因中 5 个与 GBM 有关联；综合脆弱性和中心性两个测度预测出的 15 个基因中 8 个有文献支持；而综合度和中心性两个测度预测的 11 个基因中，有 3 个得到文献证实。

表 2 三种网络测度预测 GBM 基因的结果

Table 2 The result of three measures on predicting GBM genes

基因名称	DCV	DV	CV	DC	V	D	CC	文献编号
CALR	√	---	√	√	---	---	√	---
GNAS	√	√	√	---	√	---	---	17440062
PLCG1	√	√	---	√	---	√	---	---
RPN2	√	---	√	√	---	---	√	---
CTNNB1	√	√	√	---	√	---	---	21424125、21321483
INCENP	√	---	√	√	---	---	√	---
PTGS2	√	√	√	---	√	---	---	21360625
CUL3	√	---	√	√	---	---	√	---
EIF4A1	√	---	√	√	---	---	√	---
MYC	√	√	√	---	√	---	---	19706761
RABGGTB	√	---	√	√	---	---	√	---
AURKB	√	---	√	√	---	---	√	19139420
GLI1	√	√	√	---	√	---	---	17628016
PAFAH1B1	√	---	√	√	---	---	√	20084519
RPN1	√	---	√	√	---	---	√	---
RBX1	√	---	√	√	---	---	√	19509229、14712485

Note: DC：D+CC; DV:D+V; DCV:D+CC+V; CV: C+V

上述研究结果表明，采用脆弱性结合共功能测度预测的候选 GBM 基因正确率（P_New）为 100%，结合脆弱性和度测度预测的正确率为 83.33%。令我们意外的是，综合三个测度预测的正确率仅为 50.00%，利用节点的脆弱性和中心性测度预测正确率为 53.33%，其它测度预测结果正确率均不超过 50%，中心性结合度测度的预测正确率只有 27.27%，而只用度作为测度预测出的 1 个基因也不是 GBM 相关基因，这说明节点的中心性和度测度并不能正确的预测出候选 GBM 相关基因。

2.6 综合分析

本文使用不同的网络测度结合共功能基因预测 GBM 相关基因。为了综合评价每种测度组合的效能，本研究用调和平均值度量每个测度的综合得分，分别为 61.64%，86.46%，60.96%，40.72%，93.90%，0，和 13.95%。

$$P=\frac{2\times P\_origianl\times P\_New}{P\_origianl+P\_New}$$

可以看出，利用网络的脆弱性作为测度预测候选 GBM 相关基因表现出了最好的效果。而节点的度只能作为评价已有高度节点是否为 GBM 相关基因的测度，不能够用来预测候选 GBM 相关基因。节点的中心性测度预测的结果最差，因此本文认为它不可以作为预测候选 GBM 基因的测度。尽管节点的脆弱性结合其他测度预测的结果也相对较好，但是他们都把使用网络的脆弱性预测基因的效能降低了。由以上结果可知，结合 GBM 基因网络中节点的脆弱性和共功能两个测度能够很好的预测 GBM 相关基因。

3 讨论

本文结合 GBM 基因的功能基因和来自基因关系网络中基因的测度来预测新的 GBM 相关基因。不同的网络测度预测的结果不尽相同，本节将对不同的测度进行综合的讨论。研究表明，基因关系网络中度高的节点更加倾向于与疾病相关。本文的结果也证实了这一点，即度高并且与已知 GBM 相关基因共功能的基因中有 97.78%与 GBM 有关。然而，将度高节点预测为候选 GBM 基因的结果并不好。其中的原因主要是基因网络的获得来自于 GBM 相关基因。当度与其它测度结合预测 GBM 基因时，预测的结果比其它测度单独预测效能要低，且效果不明显。从这点可以说明，节点的度不适合用来预测候选疾病基因。

用节点的脆弱性做测度也是预测候选疾病相关基因的有效方法。在前人研究中[17]，他们选取了脆弱性得分前 50 的基因作为候选基因，但没有明确给出这 50 个基因的脆弱性得分是否为正数。节点的脆弱性是指排除该节点后，网络的全网效率较低的比率。如果排除节点后，网络的全网效率增高了，说明该节点本身对网络的效率没有积极的贡献。反之，如果删除某节点使全网效率降低，说明该节点对网络有正贡献。因此，本文选取了所有脆弱性为正数的节点，将其作为候选 GBM 基因。研究结果表明，该方法预测出的 5 个基因都是与 GBM 相关的。

节点的中心性可以用来度量某节点做为中心节点的得分。有研究[17]用其作为预测新的疾病基因的测度。然而,他们并没有对这个测度和其它几种测度单独使用时的预测效能进行评价,也没有考虑预测结果是否对预测的概率产生影响。为了衡量节点的中心性和其它测度是否能够单独预测 GBM 相关基因，本文分别采用不同测度及测度的组合，结合与已知 GBM 相关基因共功能的基因进行预测，并对结果进行了比较和验证。结果发现,节点的中心性在预测新的 GBM 基因的效能上很差,并明显低于利用脆弱性预测的效能。表明节点的中心性并不能够作为预测 GBM 基因的测度,同样也说明了中心性得分高的基因预测为疾病相关基因的可能性也较低。

本文在评价预测结果的效能时,采取了识别已知 GBM 相关基因的效率，以及预测新的 GBM 基因的效能 2 个测度,并综合评价了不同测度的预测结果。识别已知 GBM 基因的效率就是计算已知基因集合在候选基因集合中的比例,这也能够作为已知基因属性的评价测度。通过衡量预测新基因的效能可以判断各测度预测出新的 GBM 基因是否为 GBM 基因。各种测度不同组合,共预测出了 16 个候选 GBM 相关基因,其中 8 个基因有文献证实,包括网络节点的脆弱性为正的 5 个基因。这说明本文应用的脆弱性测度结合共功能的方法在预测疾病基因方面有良好的效能。

参 考 文 献(References)

[1] Wuchty S, Vazquez A, Bozdag S, et al. Genome-wide associations of signaling pathways in glioblastoma multiforme [J]. BMC Medical Genomics, 2013, 6(1): 11

[2] Smardova J, Liskova K, Ravcukova B, et al. High Frequency of Temperature-Sensitive Mutants of p53 in Glioblastoma[J]. Pathology & Oncology Research, 2013: 1-8

[3] Xu Q, Yuan X, Liu G, et al. Hedgehog signaling regulates brain tumor-initiating cell proliferation and portends shorter survival for patients with PTEN-coexpressing glioblastomas[J]. Stem Cells, 2008, 26(12):3018-3026

[4] Jin G, Cook S, Cui B, et al. HDMX regulates p53 activity and confers chemoresistance to 3-bis (2-chloroethyl)-1-nitrosourea [J]. Neuro Oncol, 2010, 12(9):956-966

[5] Wiedemeyer WR, Dunn IF, Quayle SN, et al. Pattern of retinoblastoma pathway inactivation dictates response to CDK4/6 inhibition in GBM [J]. Proc Natl Acad Sci U S A, 2010, 107(25):11501-11506

[6] Liu Y, Shete S, Hosking FJ, et al. New insights into susceptibility to glioma[J]. Arch Neurol, 2010, 67(3):275-278

[7] Boettner B. Cancer: Approaching GBM via EphA3 [J]. Nature Medicine, 2013, 19(3): 278-279

[8] Nadkarni A, Shrivastav M, Mladek A C, et al. ATM inhibitor KU-55933 increases the TMZ responsiveness of only inherently TMZ sensitive GBM cells [J]. Journal of neuro-oncology, 2012, 110(3): 349-357

[9] Weng HY, Hsu MJ, Wang CC, et al. Zerumbone suppresses IKKalpha, Akt, and FOXO1 activation, resulting in apoptosis of GBM 8401 cells [J]. Biomed Sci, 2012,19(1): 86

[10] Wakimoto H, Kesari S, Farrell CJ, et al. Human glioblastoma-derived cancer stem cells: establishment of invasive glioma models and treatment with oncolytic herpes simplex virus vectors[J]. Cancer Res, 2009, 69(8):3472-3481

[11] Zhao Y, Xiao A, diPierro CG, et al. An extensive invasive intracranial human glioblastoma xenograft model: role of high level matrix metalloproteinase 9[J]. Am J Pathol, 2010, 176(6):3032-3049

[12] Ma XJ, Yin HJ, Chen KJ. Differential gene expression profiles in coronary heart disease patients of blood stasis syndrome in traditional Chinese medicine and clinical role of target gene [J]. Chin J Integr Med, 2009, 15(2):101-106

[13] Taylor BS, Barretina J, Socci ND, et al. Functional copy-number alterations in cancer[J]. PLoS One, 2008, 3(9):e3179

[14] Latora V, Marchiori M. Efficient behavior of small-world networks [J]. Phys Rev Lett, 2001, 87(19):198701

[15] Ames G M, George D B, Hampson C P, et al. Using network properties to predict disease dynamics on human contact networks[J]. Proceedings of the Royal Society B: Biological Sciences, 2011, 278 (1724): 3544-3550

[16] Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network [J]. BMC Genomics, 2010, 11 Suppl 3:S5

[17] Ortutay C, Vihinen M. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies[J]. Nucleic Acids Res, 2009, 37 (2):622-628

[18] Dipple KM, Phelan JK, McCabe ER. Consequences of complexity within biological networks: robustness and health, or vulnerability and disease[J]. Mol Genet Metab, 2001, 74(1-2):45-50

[19] Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms [J]. Bioinformatics ,2010, 26(18):i561-567

[20] Guan Y, Ackert-Bicknell CL, Kell B, et al. Functional genomics complements quantitative genetics in identifying disease-gene associations[J]. PLoS Comput Biol, 2010, 6(11):e1000991