PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(3): 286~296
www.pibb.ac.cn

基于全局角度网络策略的复杂疾病风险通路识别 *

邓莉莉 许艳军 张春龙 姚茜岚 冯 丽 李春权 **
(哈尔滨医科大学生物信息科学与技术学院，哈尔滨 150081)

摘要 复杂疾病的发生发展与机体内生物学通路的功能紊乱有密切联系，从高通量数据出发，利用计算机辅助方法来研究疾病与通路间的关系具有重要意义．本文提出了一个新的基于网络的全局性通路识别方法．该方法利用蛋白质互作信息和通路的基因集组成信息构建复杂的蛋白质 - 通路网．然后，基于表达谱数据，通过随机游走算法从全局层面优化疾病风险通路．最终，通过扰动方式识别统计学显著的风险通路．将该网络运用于结肠直肠癌风险通路识别，识别出 15 个与结肠直肠癌发生与发展过程显著相关的通路．通过与其他通路识别方法(超几何检验，SPIA)相比较，该方法能够更有效识别出疾病相关的风险通路．

关键词 风险通路，蛋白质 - 通路网，蛋白质互作，随机游走
学科分类号 R318.04，Q811.4，R735.3 **DOI**: 10.16476/j.pibb.2014.0105

复杂疾病的发病机理被认为是来自于多基因变异的联合效应[1]，疾病的发生和发展与生物通路功能紊乱有密切联系[2]，此外药物治疗中也涉及代谢通路的局部改变[3]．随着高通量技术的发展，产生了海量的生物学数据．因此，从高通量数据出发，利用生物信息学方法挖掘与疾病发生机制相关的通路，对疾病的诊断和治疗具有重要意义[4]．

生物通路主要由基因产物构成，传统的通路分析被简化为基因集的富集分析[5]．目前通用的分析方法有两大类，分别是过表达分析(over-representation analysis，ORA)和功能类得分(functional class scoring，FCS)[6]．ORA 方法主要思想是将感兴趣的基因注释到通路中，用卡方检验、超几何检验等方法计算每个通路被富集的显著性．虽然 ORA 方法已经被广泛应用，但该类方法仍有一些缺陷，这些检验方法只考虑了基因的数目，忽略了基因间的区别．FCS 方法是基于 ORA 方法的改进，该类方法用到了通路中的所有基因，利用基因表达信息对单个通路中的每个基因计算一个统计量(如 anova，*t*-test，*z*-score)并通过加和等方法将其转化成通路统计量，评估通路显著性．通路中的基因是通过化学互作、信号转导等相互作用来执行功能的，通路并不等同于基因集，FCS 方法也没有考虑到通路中基因与基因间的相互作用，因此 Fang 等[7]引进了网络分析，利用基因功能关系网来计算每个基因的权重，进而得到通路统计量，提高了疾病风险通路的识别能力．Liu 等[8]利用有向随机游走(DRW)方法，考虑每个基因在通路结构中的拓扑重要性，提高了通路对疾病分类的准确性．其实从全局角度来看，单个通路是大的生物互作网络的一部分，许多关键基因能在多个通路中发挥作用，因此通路与通路间有着直接或间接的功能联系．但是，目前大多数通路识别方法只是独立分析单个通路，只考虑通路内部的基因．一个通路的异常可能导致相关通路的紊乱，在通路分析中不应该忽视通路间的联系，从全局角度来分析与识别疾病风险通路是很有必要的．

在本文中我们提出了一个新的基于网络的全局性通路识别方法，利用蛋白质互作信息和通路的基

* 国家自然科学基金(31200996), 黑龙江省教育厅项目(12531295)和哈尔滨医科大学于维汉院士杰出青年培养基金资助．
** 通讯联系人．
Tel: 13704807434, E-mail: lcqbio@aliyun.com
收稿日期：2014-10-29，接受日期：2015-01-15

因集组成信息构建复杂的蛋白质 - 通路交互网，基于表达谱数据从全局层面优化疾病风险通路．本文构建的蛋白质 - 通路交互网加强了蛋白质间的相互作用和通路间的联系．在随机游走过程中考虑了基因的差异表达量，弥补了传统方法的缺陷．我们将该方法运用到结肠直肠癌的风险通路识别中，与传统方法相比，有效地识别出与该疾病有显著关系的风险通路，对于今后研究疾病的发生发展机制有重要的指导意义．

1　材料与方法

1.1　数据

1.1.1　差异表达基因数据

利用 GEO(Gene Expression Omnibus，http://www.ncbi.nlm.nih.gov/geo/)中的结肠直肠癌表达谱数据(GSE8671 和 GSE9348)，其中 GSE8671 数据检测了 32 个结肠癌腺瘤组织和 32 个配对的癌旁组织，GSE9348 数据中检测了 70 个结肠直肠癌早期患者的肿瘤组织和 12 个健康人的组织．用倍数分析方法识别差异表达基因，通过设置倍数分析方法的阈值(FC>2 或 FC<–2)，在 GSE8671 表达谱数据中从 20 184 个基因中识别出 1 859 个差异基因，在 GSE9348 表达谱中从 20 184 个基因中识别出 2 388 个差异基因．

1.1.2　蛋白质互作数据

蛋白质互作数据来源于人类蛋白质参考数据库(HPRD)[9](http://www.hprd.org/)，从 HPRD 中下载蛋白质互作数据后，对互作数据进行网络构建前的预处理，去除有缺失值的互作对，以及合并冗余的互作对，最终得到 36 867 个蛋白质 - 蛋白质互作关系对．

1.1.3　通路数据

通路数据来源于 KEGG(Kyoto Encyclopedia of Genes and Genomes，http://www.genome.jp/kegg/)，其中的通路数据库(PATHWAY database)存储了每个通路的基因信息，本文使用了 PATHWAY database 中的 300 个通路，其中包括 150 个代谢通路和 150 个非代谢通路．我们利用课题组先前开发的 iSubpathwayMiner[10-11] R 软件包将 KEGG 中每个通路进行图形转化，由于一些通路里并不包含蛋白质或编码蛋白质的人类基因，最终分别获得了构成 85 个代谢通路和 134 个非代谢通路的蛋白质集合，并把这 219 个通路作为识别疾病风险通路的候选．

1.2　方法

首先从 HPRD 和 KEGG 数据库中提取蛋白质互作信息和通路的组成信息，构建蛋白质 - 通路交互网，然后把感兴趣的差异基因作为种子节点注释到网络中，利用随机游走算法对其中通路节点分配权重，并通过网络扰动来评价候选通路的显著性，最终通过显著性阈值识别出疾病风险通路．方法流程如图 1 所示．

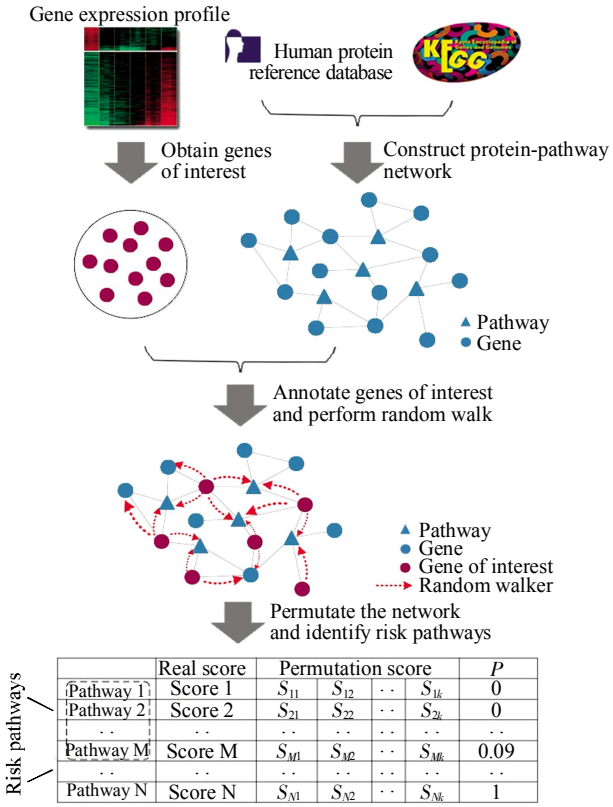

Fig. 1　Flow diagram of the methodology

1.2.1　整合通路和蛋白质互作信息构建蛋白质 - 通路交互网

把从 HPRD 中提取的人类蛋白质互作对以及从 KEGG 中获取的通路和通路中蛋白质的对应关系作为构建整合网络的基础．首先利用从 HPRD 中提取的蛋白互作对构建蛋白质互作网络，把所有有互作关系的蛋白对连接起来，然后把从 KEGG 中获得的候选通路作为节点整合进网络，如果蛋白质互作网中某一蛋白质在一通路中出现，那么将此通路与该蛋白质连接起来．这样便构建出一个整合了蛋白质互作和通路构成信息的蛋白质 - 通路交互网．

1.2.2　基于网络利用 RW 方法对通路进行优化

基于网络的随机游走(random walk，RW)算法用来表示一种不规则的变动形式，网络中一些感兴趣的节点被赋予权值作为种子节点，从种子节点开始沿网络结构向其他邻居节点分配权值，同时在游走过程中也会得到其他节点权值的分配，最终使得与种子节点联系紧密的节点倾向于有更高的权重. 这里使用的是重启型随机游走(random walk with restart，RWR)，公式定义为

$$p^{t+1}=(1-r)Wp^{t}+rp^{0} \quad (1)$$

其中 r 表示某一节点在每一次游走中将其权值分配给邻居节点的概率，这里使用默认值 0.7，W 表示网络的标准化邻接矩阵，p^0 表示节点的初始权重向量，p^t 表示网络 t 次游走后节点新权重.

利用重启型随机游走算法给通路打分的具体方法如图 2 所示. 图 2a 所示为用来进行随机游走的网络示意图，其中 2，4 代表种子节点. 图 2b 中矩阵是图 2a 所示网络图的标准化邻接矩阵，即公式(1)中的 W. 图 2c 中向量的每个元素都分别对应图 2a 中每个节点的初始权重，即公式(1)中的 p^0. 图 2d 表示对图 2a 中网络根据公式(1)进行一次随机游走后得到的节点新权重，图 2e 所示为经过多次随机游走直到 $|p^{t+1}-p^{t}|<10^{-10}$ 时得到的新权重状态，节点颜色越深表示得分越高，其中 p^* 表示节点的最终得分. 在本文中将疾病差异表达基因作为种子节点，把差异基因的 fold change 值标准化后作为种子节点的初始权值，网络中其他节点的权值设为 0. 当 $|p^{t+1}-p^{t}|<10^{-10}$ 终止迭代，把 p^* 作为最终的节点得分. 通路节点得分越高，认为与差异基因联系越紧密，则越倾向于作为疾病风险通路.

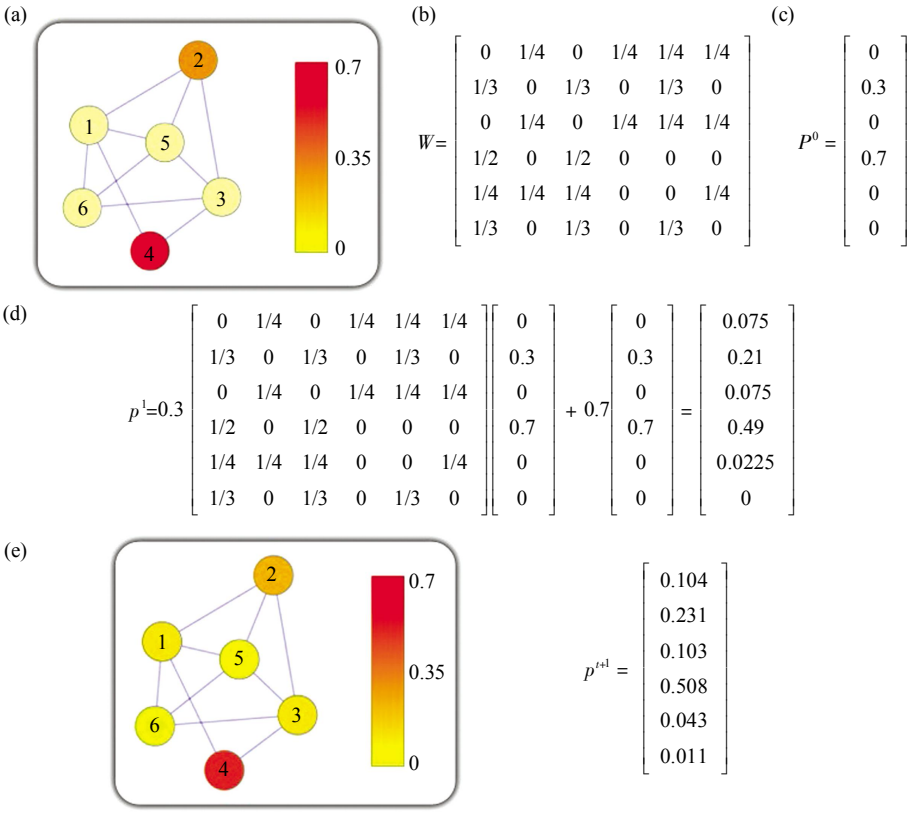

Fig. 2　The process of the random walk

(a) The network is used to execute random walk and the different colors in nodes represent different weights. (b) The standard adjacent matrix of the network in subgraph (a). (c) The initial weights of the nodes in subgraph (a). (d) The new weights of the nodes in subgraph (a) after the first step of random walk. (e) The final network state after random walk and the vector represents the weights of the nodes in the new graph.

1.2.3　扰动网络评估通路显著性

KEGG 中有一些通路与特定疾病并没有密切联系，但因为通路规模较大(通路中包含的节点多)，在我们构建的蛋白质 - 通路交互网中有很高的度，因此进行网络随机游走时倾向于获得更高的得分，这样就可能导致识别结果具有一定的偏性. 为了消

除这样的偏性，我们做随机化扰动，扰乱种子节点在网络中的分布，其中网络的具体扰动方法为从真实网络中随机选取和识别出的疾病差异表达基因个数相等的节点作为种子节点，然后把差异基因的 fold change 值随机分配给这些种子节点作为权重，进行随机游走，计算各通路的得分．按如上方法对网络进行 1000 次扰动后，统计每一个通路的随机得分中高于真实得分的次数 N，计算候选通路的显著性 $P=N/1000$．对于那些在蛋白质 - 通路交互网中有很高的度的通路节点来说，在随机扰动的过程中仍然有大量与其紧密相连的节点为其分配权值，使得相应的 P 值会增高．最后通过阈值 $P<0.01$，我们能够校正通路规模对结果的影响，识别出疾病显著相关通路．最后用 P 值来评价风险通路与疾病的显著关系．

2 结　　果

我们将本文提出的基于网络的全局通路识别方法应用到结肠直肠癌表达谱数据(GEO：GSE8671)中，利用全局的随机游走对通路进行优化，并扰动网络获得最终的 P 值．表 1 显示了 $P<0.01$ 的 15 个结肠直肠癌风险通路．

为了检验本文方法的有效性，我们与经典的通路识别方法(超几何检验)相比较．结果显示，超几何检验方法识别出了 15 个结肠直肠癌风险通路．

Table 1　The risk pathways identified by our method

Pathway ID	Pathway name	Annotation	*P*
path:04062	Chemokine signaling pathway	39	0
path:04514	Cell adhesion molecules (CAMs)	30	0
path:04972	Pancreatic secretion	21	0
path:04610	Complement and coagulation cascades	22	0
path:04060	Cytokine-cytokine receptor interaction	60	0.001
path:00910	Nitrogen metabolism	8	0.001
path:04640	Hematopoietic cell lineage	26	0.001
path:04964	Proximal tubule bicarbonate reclamation	7	0.001
path:04974	Protein digestion and absorption	19	0.002
path:00460	Cyanoamino acid metabolism	2	0.003
path:02010	ABC transporters	9	0.004
path:04672	Intestinal immune network for IgA production	14	0.004
path:00140	Steroid hormone biosynthesis	12	0.007
path:05150	Staphylococcus aureus infection	15	0.008
path:04670	Leukocyte transendothelial migration	19	0.009

*P < 0.01. Annotation: The number of differential genes in the corresponding pathway.

通过两种方法对比分析，发现我们的方法识别出一些新的疾病风险通路 (图 3)．两种方法对比结果如表 2 所示，表中共包含 20 个通路，这些通路为本方法与超几何检验方法识别出的显著性通路的并集．

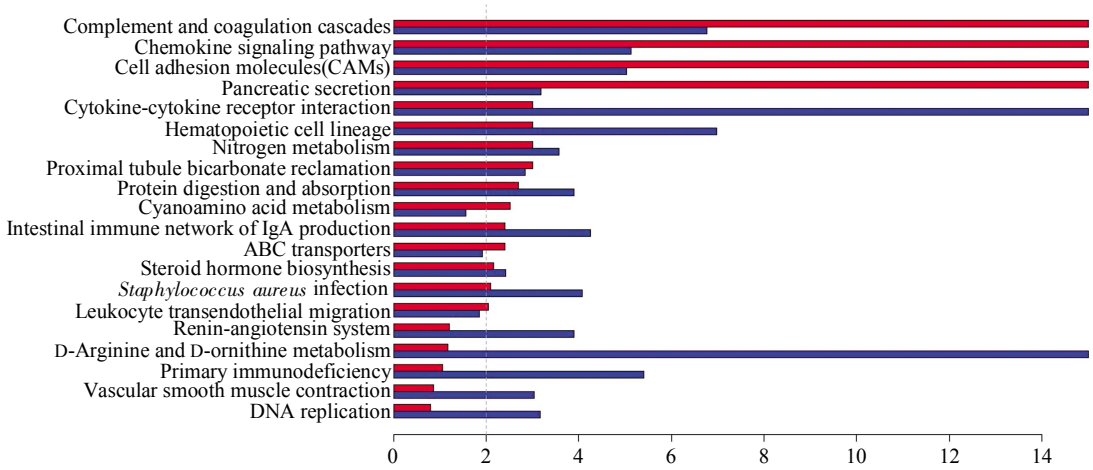

Fig. 3　The comparison of identified results from our method and the hypergeometric test

The different colors on the vertical axis represent the risk pathways are identified by different methods. The red represent the risk pathways only identified by our method, the green represent the risk pathways only identified by hypergeometric test, the blue represent the risk pathways identified by our method and hypergeometric test. The length of the bar is $L=-\lg(p)$(p represent the significant probability of the corresponding pathway, the red bars represent the pathways identified by our method, the bule bars represent the pathways identified by hypergeometric test). The grey vertical dashed line represents the length of the bar when p is equal to 0.01.

Table 2 The comparison of the risk pathways identified by our method and the hypergeometric test

Pathway ID	Pathway name	Annotation	Rank1	Rank2
path:04062	Chemokine signaling pathway	39	1	6
path:04514	Cell adhesion molecules (CAMs)	30	2	7
path:04972	Pancreatic secretion	21	3	13
path:04610	Complement and coagulation cascades	22	4	4
path:04060	Cytokine-cytokine receptor interaction	60	5	2
path:00910	Nitrogen metabolism	8	6	12
path:04640	Hematopoietic cell lineage	26	7	3
path:04964	Proximal tubule bicarbonate reclamation	7	8	17
path:04974	Protein digestion and absorption	19	9	10
path:00460	Cyanoamino acid metabolism	2	10	46
path:02010	ABC transporters	9	11	35
path:04672	Intestinal immune network for IgA production	14	12	8
path:00140	Steroid hormone biosynthesis	12	13	23
path:05150	Staphylococcus aureus infection	15	14	9
path:04670	Leukocyte transendothelial migration	19	15	39
path:00472	D-Arginine and D-ornithine metabolism	1	37	1
path:05340	Primary immunodeficiency	13	43	5
path:04614	Renin-angiotensin system	7	36	11
path:03030	DNA replication	10	59	14
path:04270	Vascular smooth muscle contraction	24	53	15

*Rank1: The significant order of the risk pathways identified by our method. Rank2: The significant order of the risk pathways identified by the hypergeometric test. Annotation: The number of differential genes in the corresponding pathway.

为了直观对比分析两种方法的结果，我们还绘制了条形图(图 3). 在该图中红色条形图表示本文方法识别出的通路，蓝色条形图表示超几何方法识别出的通路，条形图的长度 $L=-\lg(p)$，其中 p 代表两种方法计算得到的相应通路的显著性概率值. 由于我们在筛选风险通路时统一阈值为 $P<0.01$(即 $L>2$)，因此若条形图长度超过图 3 中的灰色垂直虚线，则表明该通路为风险通路. 同时在图 3 的纵轴上，用红色标示的通路名称代表只被本文方法识别出来，蓝色的表示只被超几何方法识别出来，绿色的表示被两种方法均识别出来.

从表 2 和图 3 中可以看出，新识别出来的风险通路有近端小管碳酸氢盐回收(path:04964)、ABC 转运子(path:02010)、类固醇激素生物合成(path:00140)等通路，大量研究证实这些通路跟结肠直肠癌的发生机制有一定的联系. 其中近端小管碳酸氢盐回收(path:04964)排在第 8 位. 在正常的生理活动中结肠组织扮演着重要角色，包括维生素、盐、营养物质以及水的吸收，被认为是消化过程的最后阶段[12-13]. 其中近端小管碳酸氢盐回收通路中的 ATP 激酶、Na^+/K^+ 转运子、谷氨酸脱氢酶等分子对于钠钾离子的交换是必需的，而且为不同的营养物质的主动转运提供能量[14]. 因此该通路的失调在结肠直肠癌的发生过程中起一定的作用. ABC 转运子通路排在第 11 位. 有文章表明，ABC 基因家族中的转运蛋白在肿瘤细胞的多药耐药性中扮演重要角色，该通路的失调可能导致癌症干细胞对常规治疗有抵抗力，并且导致肿瘤的再生长和复发[15]. 类固醇激素生物合成通路也被证实在一定程度上与结肠直肠癌有关. 据文献报道，患有乳腺癌的妇女会增加患结肠直肠癌的风险[16].

此外，用我们的方法同样也得到了传统方法所识别出来的风险通路，如趋化因子信号通路(path:04062)、细胞黏附分子(path:04514)、氮代谢(path:00910)，这些通路与结肠直肠癌的发生发展机制也有密切关系. 但与超几何检验方法相比，这些通路排序位置从第 6、7、12 位上升到了第 1、2、6 位. 其中排在第 1 位的是趋化因子信号通路(图 4),

该通路中的 PI3K 是一种胞内磷脂酰肌醇激酶，与 *v.src* 和 *v.ras* 等癌基因的产物相关．且 PI3K 本身具有丝氨酸 / 苏氨酸(Ser/Thr)激酶的活性，该酶的失活会导致与结肠直肠癌发生机制相关的 3 个基因功能失调．这一发现可能为癌症治疗提供新的靶点[17]．排在第 2 位的细胞黏附分子通路中的 L1-CAM，是一种神经元细胞黏附受体，在很多癌细胞中表达[18]．据文献报道，大多数结肠直肠癌都是由于一种关键蛋白 β-catenin 的变化而引发的[19]．最新研究表明，L1-CAM 作为 β-catenin 的靶基因，在人类结肠癌组织的转移过程中起到关键作用，CAMs 的异常很可能导致癌症的发生．有文献报道 80% 的结肠癌的发生跟饮食有关，大量食肉会改变体内的氮代谢，增加 NOCs 等可能的致癌物质的产生[20]．在结肠部位蛋白质的降解和代谢产物包括氮、酚类、吲哚类物质，这些物质均有毒性作用[21]．因此表明氮代谢通路也是结肠直肠癌的一个风险通路．

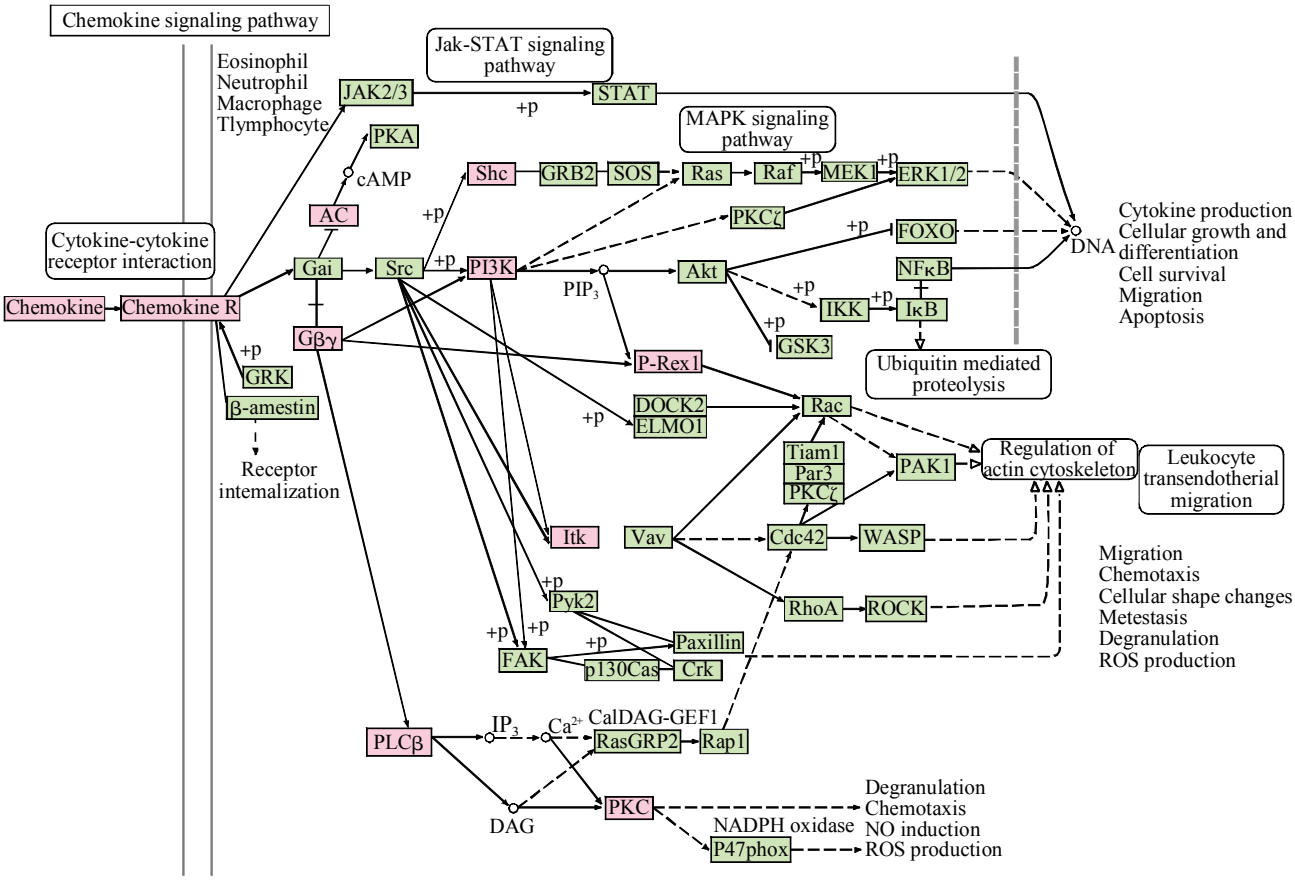

Fig. 4　Chemokine signaling pathway

The red nodes represent the differential genes annotated to this pathway, and the green nodes represent the non-differential genes.

同时，我们还将本文方法与 SPIA(signaling pathway impact analysis)方法进行了比较，该方法是一种最为流行的 DRW 策略识别通路的方法之一[22]．用 SPIA 方法处理结肠直肠癌表达谱数据(GEO：GSE8671)后，共识别出了 4 个风险通路，结果如表 3 和图 5 所示，图表中包含的通路是两种方法识别出来的显著性通路的并集．

从表 3 和图 5 可以看出，SPIA 方法识别出的 4 个风险通路也均被本文方法所识别(path:04062，path:04610，path:04060，path:04670)，同时我们的方法识别出了更多的与结肠直肠癌有关的风险通路，比如说细胞黏附分子(path:04514)、氮代谢(path:00910)等通路如前文所述，均已被证明与结肠直肠癌有密切关系[18-21]．同时还有一些没有文献证实的通路可以看做本文方法新识别出的风险通路．由此表明，我们提出的方法具有较高的准确性和有效性．

Table 3　The comparison of pathways identified by our method and the SPIA

Pathway ID	Pathway name	Method1	Method2
path:04062	Chemokine signaling pathway	Y	Y
path:04514	Cell adhesion molecules (CAMs)	Y	N
path:04972	Pancreatic secretion	Y	N
path:04610	Complement and coagulation cascades	Y	Y
path:04060	Cytokine-cytokine receptor interaction	Y	Y
path:00910	Nitrogen metabolism	Y	N
path:04640	Hematopoietic cell lineage	Y	N
path:04964	Proximal tubule bicarbonate reclamation	Y	N
path:04974	Protein digestion and absorption	Y	N
path:00460	Cyanoamino acid metabolism	Y	N
path:02010	ABC transporters	Y	N
path:04672	Intestinal immune network for IgA production	Y	N
path:00140	Steroid hormone biosynthesis	Y	N
path:05150	Staphylococcus aureus infection	Y	N
path:04670	Leukocyte transendothelial migration	Y	Y

*Method1,Method2 represent our method and the SPIA,respectively; Y represents the pathway was identified by the corresponding method; N represents the pathway was not identified by the corresponding method.

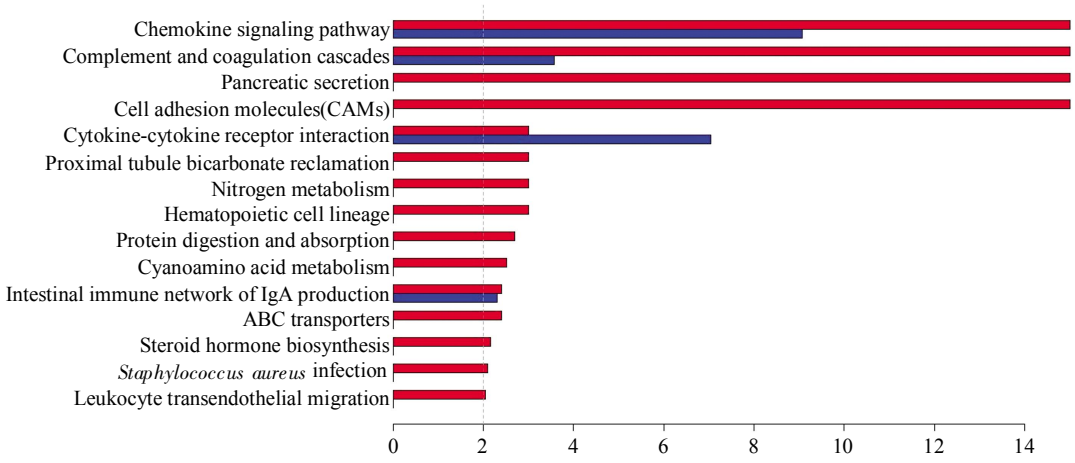

Fig. 5　The comparison of identified results from our method and the SPIA

The different colors on the vertical axis represent the risk pathways are identified by different methods. The red represent the risk pathways only identified by our method, the green represent the risk pathways only identified by SPIA, the blue represent the risk pathways identified by our method and SPIA. The length of the bar is $L=-\lg(p)$(p represent the significant probability of the corresponding pathway, the red bars represent the pathways identified by our method, the bule bars represent the pathways identified by SPIA). The grey vertical dashed line represents the length of the bar when p is equal to 0.01.

为了客观评价本方法的准确性和有效性，我们另外选取了一套结肠直肠癌表达谱数据(GEO：GSE9348)，将其运用到基于网络的全局通路识别方法中．同样通过设定阈值 $P<0.01$，共识别出了7个风险通路，结果如表4所示.

在表4列出的7个风险通路中，有3个通路(path:04060，path:04062，path:04964)是我们在运用第一套结肠直肠癌表达谱数据(GEO：GSE8671)时同样也识别出来的．其中趋化因子信号通路(path:04062)和近端小管碳酸氢盐回收(path:04964)如前文所述，已经有文章表明与结肠直肠癌有密切关系[12-14, 17]．利用该套数据识别出的其他通路也跟

疾病有一定关系，如黏着斑通路(path:04510)、癌症通路(path:05200)．在黏着斑通路(path:04510)中的蛋白酪氨酸激酶(p125FAK)在信号通路中可调节细胞黏附、迁移和生长．有文章表明，与正常黏膜相比，在结肠直肠癌肿瘤转移中 p125FAK 的表达有明显升高，p125FAK 的过表达可以被看作是出现在结肠直肠癌细胞中的一种生物标记[23]．

Table 4　The risk pathways identified by our approach with the dataset GSE9348

Pathway ID	Pathway name	*P*
path:04060	Cytokine-cytokine receptor interaction	0
path:04062	Chemokine signaling pathway	0
path:04510	Focal adhesion	0
path:05200	Pathways in cancer	0
path:05222	Small cell lung cancer	0.002
path:05146	Amoebiasis	0.005
path:04964	Proximal tubule bicarbonate reclamation	0.009

*P < 0.01.

我们把利用本文方法从 GSE8671 数据中识别出的 15 个风险通路和利用 GSE9348 数据识别出的 7 个风险通路取交集(图 6)，以全部通路(219 个)为背景，做超几何检验，公式为

$$P=1-\sum_{x=0}^{x-1}\frac{\binom{t}{x}\binom{m-t}{n-x}}{\binom{m}{n}}$$

其中 m 表示全部通路的个数(219 个)，n 表示利用 GSE9348 数据识别出的风险通路个数(7 个)，

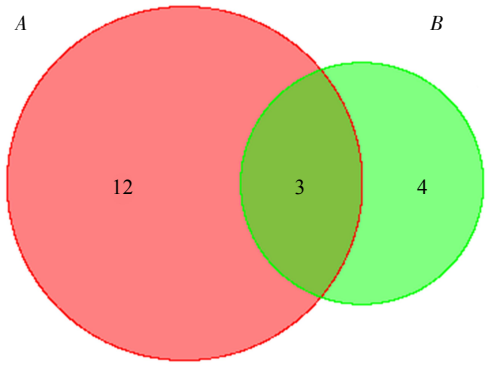

Fig. 6　The Venn diagram of the identified results by using data GSE8671 and GSE9348

A: The number of risk pathways identified by using data GSE8671.
B: The number of risk pathways identified by using data GSE9348.

t 表示利用 GSE8671 数据识别出的风险通路个数(15 个)，r 表示两套数据识别出的风险通路的交集(3 个)．最后计算得到 P=0.00 045(P < 0.01)，由此表明利用我们的方法从两套结肠直肠癌数据中识别出的风险通路是显著相关的，这种基于网络的全局通路识别方法是可靠的．

为了更直观地展示蛋白质 - 通路交互网络在疾病风险通路识别过程中的全局性优势，我们以每个识别出来的通路节点为中心向外扩一步邻居节点得到子图 a(图 7a)．从图中可以看出用本文方法识别出来的风险通路通过共享的蛋白质被紧密地联系起来了，表明疾病风险通路间是有相互作用的，并不是独立行使其功能．在开发通路识别方法时应该关注通路间的相互联系，从全局角度识别风险通路．进一步，我们以细胞黏附分子(path:04514)和白血球内皮细胞迁移通路(path:04670)为中心向外扩一步邻居节点得到子图 b(图 7b，两通路间共享 35 个蛋白)，以趋化因子信号通路(path:04062)和白血球内皮细胞迁移通路为中心向外扩一步邻居节点得到子图 c(图 7c，两通路间共享 31 个蛋白)．其中细胞黏附分子和趋化因子信号通路已被证实与结肠直肠癌的发生有关联，这两个通路的异常很可能导致与其有紧密联系的白血球内皮细胞迁移通路功能的紊乱，事实上我们的方法也识别出了该通路与疾病的发生有关联．上述结果表明，在通路识别中从全局角度出发考虑通路间的联系具有重要意义，能够识别出潜在新的疾病风险通路，提高方法的准确性．

但是与传统方法(如超几何检验)相比也有一些通路没有被本文方法识别出来，如精氨酸和鸟氨酸代谢(path:00472)、肾素 - 血管紧张素系统(path:04614)、DNA 复制(path:03030)等通路．原因可能是通路结构太小，在蛋白质 - 通路交互网中度很小，而且包含的差异基因数目过少，不足以被显著地识别出来．总体来说，通过与传统的通路识别方法进行比较以及文献证实，本文提出的从全局角度出发基于网络优化风险通路的分析方法，能够有效地识别出疾病风险通路．

3　结　　论

传统的通路识别方法一般都是单一地分析每个通路，侧重考虑通路某一方面的重要性，如考虑差异基因的注释数目、基因的权重等，因而总有一些通路被这些方法忽略掉，而这些通路通常被证明与疾病的产生机制有密切联系．因此，在本研究中我

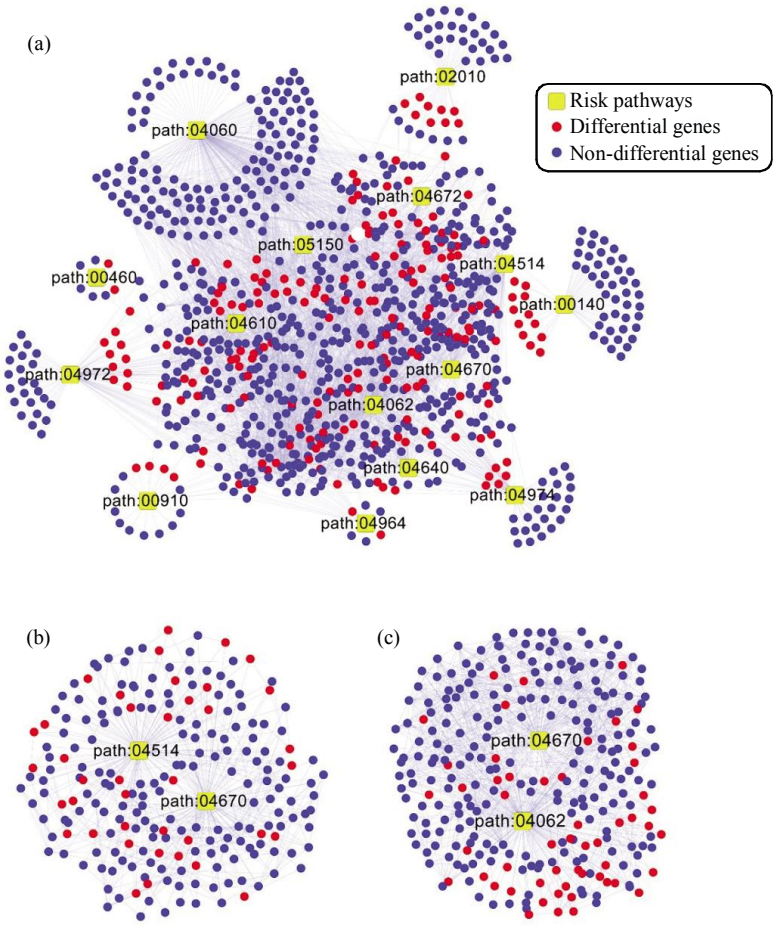

Fig. 7 Some subgraphs composed of risk pathways and their first neighbor nodes

(a) The subgraph composed of risk pathways identified by our method and their first neighbor nodes. (b) The subgraph composed of cell adhesion molecules (CAMs) pathway, leukocyte transendothelial migration pathway and their first neighbor nodes. (c) The subgraph composed of Chemokine signaling pathway, leukocyte transendothelial migration pathway and their first neighbor nodes.

们利用构建的蛋白质 - 通路交互网和随机游走策略，把通路内部和外部基因间的相互联系，通路与通路间的联系以及基因在行使生物学功能时的差异考虑到了通路的优化过程中，提出的基于网络的观点从全局角度筛选疾病风险通路的方法具有较高的准确性和有效性，为识别和分析疾病风险通路提供了一种新思路.

参 考 文 献

[1] Li C, Han J, Shang D, *et al*. Identifying disease related sub-pathways for analysis of genome-wide association studies. Gene, 2012, **503**(1): 101−109

[2] Li X, Li C, Shang D, *et al*. The implications of relationships between human diseases and metabolic subpathways. PloS One, 2011, **6**(6): e21131

[3] Li C, Shang D, Wang Y, *et al*. Characterizing the network of drugs and their affected metabolic subpathways. PloS One, 2012, **7**(10): e47326

[4] Haynes W A, Higdon R, Stanberry L, *et al*. Differential expression analysis for pathways. PLoS Computational Biology, 2013, **9**(3): e1002967

[5] Gu Z, Liu J, Cao K, *et al*. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Systems Biology, 2012, **6**(1): 56

[6] Khatri P, Sirota M, Butte A J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Computational Biology, 2012, **8**(2): e1002375

[7] Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell Research, 2011, **22**(3): 565−580

[8] Liu W, Li C, Xu Y, *et al*. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 2013, **29**(17): 2169−2177

[9] Prasad T K, Goel R, Kandasamy K, *et al*. Human protein reference database——2009 update. Nucleic Acids Research, 2009, **37**(suppl 1): D767−D772

[10] Li C, Li X, Miao Y, *et al*. SubpathwayMiner: a software package for flexible identification of pathways. Nucleic Acids Research, 2009, **37**(19): e131−e131

[11] Li C, Han J, Yao Q, *et al*. Subpathway-GM: identification of metabolic subpathways *via* joint power of interesting genes and metabolites and their topologies within pathways. Nucleic Acids Research, 2013, **41**(9): e101−e101

[12] Weiser M M, Bloor J H, Dasmahapatra A. Intestinal calcium absorption and vitamin D metabolism. Journal of Clinical Gastroenterology, 1982, **4**(1): 75−86

[13] Rose R C. Water-soluble vitamin absorption in intestine. Annu Rev Physiol, 1980, **42**(1): 157−171

[14] Magdeldin S, Yoshida Y, Li H, *et al*. Murine colon proteome and characterization of the protein pathways. BioData Mining, 2012, **5**(1): 11

[15] Leonard G D, Fojo T, Bates S E. The role of ABC transporters in clinical practice. The Oncologist, 2003, **8**(5): 411−424

[16] Leo A, Messa C, Cavallini A, *et al*. Estrogens and colorectal cancer. Current Drug Targets-Immune, Endocrine & Metabolic Disorders, 2001, **1**(1): 1−12

[17] Parsons D W, Wang T-L, Samuels Y, *et al*. Colorectal cancer: mutations in a signalling pathway. Nature, 2005, **436**(7052): 792−792

[18] Gavert N, Sheffer M, Raveh S, *et al*. Expression of L1-CAM and ADAM10 in human colon cancer cells induces metastasis. Cancer research, 2007, **67**(16): 7703−7712

[19] Chung G G, Provost E, Kielhorn E P, *et al*. Tissue microarray analysis of β-catenin in colorectal cancer shows nuclear phospho-β-catenin is associated with a better prognosis. Clinical Cancer Research, 2001, **7**(12): 4013−4020

[20] Bingham S. Diet and colorectal cancer prevention. Biochemical Society Transactions, 2000, **28**(2): 12−16

[21] Hughes R, Magee E, Bingham S. Protein degradation in the large intestine: relevance to colorectal cancer. Current Issues in Intestinal Microbiology, 2000, **1**(2): 51−58

[22] Tarca A L, Draghici S, Khatri P, *et al*. A novel signaling pathway impact analysis. Bioinformatics, 2009, **25**(1): 75−82

[23] Han N M, Fleming R D, Curley S A, *et al*. Overexpression of focal adhesion kinase (p125FAK) in human colorectal carcinoma liver metastases: independence from c-src or c-yes activation. Annals of Surgical Oncology, 1997, **4**(3): 264−268

A Network-based Strategy From The Global Perspective for Identification of Risk Pathways in Complex Diseases*

DENG Li-Li, XU Yan-Jun, ZHANG Chun-Long, YAO Qian-Lan, Feng Li, LI Chun-Quan**

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

Abstract The initiation and progression of complex diseases have a close relationship with dysfunction of biological pathways in our body. Developing computational techniques to study the relationship between diseases and pathways through high-throughput data has essential biological significance. However, the traditional identification approaches of pathways which are significantly related to experiment conditions usually reduce pathways to gene sets. It is obvious that these methods do not consider the interactions between genes and the different roles that genes play in pathways, and they don't fully mine pathway information. Therefore we integrated protein-protein interaction information and gene weights into pathway analysis, and constructed a protein-pathway network which contains information in protein-protein interactions and pathways. We then scored pathways by random walk algorithm to optimize disease risk pathways. Finally, the statistically significant pathway can be identified through permutation method. We applied the network to a colorectal cancer dataset, and finally identified fifteen pathways which are significantly related to the development of this disease. Compared with other pathway identification methods (hypergeometric test and SPIA), our approach can effectively identify risk pathways related to complex diseases. In order to test the stability of our method in identifying risk pathways related to diseases, we used our method to identify risk pathways by using another colorectal cancer dataset. We found that the identified results can prove the stability of our method.

Key words risk pathways, protein-pathway network, protein-protein interactions, random walk

DOI: 10.16476/j.pibb.2014.0105

*This work was supported by grants from The National Natural Science Foundation of China (31200996), The Education Department Project of Heilongjiang Province (12531295), and Yu Weihan Outstanding Youth Training Fund of Harbin Medical University.

**Corresponding author.

Tel: 86-13704807434, E-mail: lcqbio@aliyun.com

Received: October 29, 2014 Accepted: January 15, 2015