PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(2): 176~186
www.pibb.ac.cn

基于扩展起始节点和加权融合策略预测肺癌风险致病基因 *

王一斌 程咏梅 张绍武 **
(西北工业大学自动化学院，信息融合技术教育部重点实验室，西安 710072)

摘要 肺癌风险致病基因预测有助于了解疾病发病机制、提高临床治疗效果．目前，以重启游走为框架的风险致病基因预测算法，普遍存在起始节点少、节点转移概率相同、信息源单一的问题．为此，本文提出一种基于扩展起始节点和加权融合策略的风险致病基因预测算法(命名为 AFMFSC)，并在肺癌中验证算法有效性．首先，基于增广模糊测量思想，计算疾病表型近似基因间的增广功能相似得分，从中选出重要基因与致病基因作为扩展起始节点；其次，采用节点拓扑相似度转移矩阵及基因表达差异相关性转移矩阵，分别在蛋白质网络中重启随机游走，并将两种结果加权融合排序；最后，通过富集分析排名靠前基因，得到有显著意义的风险致病基因．AFMFSC 算法预测的 73 个肺癌风险致病基因，均与肺癌发生、发展有密切联系，生物学意义显著．与其他排序算法相比，AFMFSC 算法的 Top 1%、Top 5%和 AUC 值比较大，平均排名和受拓扑特性偏差影响程度小；融合策略排名性能优于单一转移矩阵或普通邻接矩阵游走排名．AFMFSC 算法不仅能准确有效地预测肺癌风险致病基因，而且可推广预测其他疾病风险致病基因，为探索癌症致病机理提供新视角及依据．

关键词 风险致病基因，扩展起始节点，拓扑相似度转移矩阵，基因表达差异相关性转移矩阵，重启随机游走
学科分类号 R318.04，Q78 **DOI**: 10.16476/j.pibb.2015.0380

在人类癌症疾病中，最容易确诊且死亡率最高的是肺癌．它是一种细胞失控生长且由多个基因共同作用导致的复杂疾病．近 5 年来，该疾病在世界范围内死亡率高达 80%～85%，其发病率和死亡率在所有恶性肿瘤中占男性的第一位，女性的第二位[1-3]．肺癌的病因和机理至今尚不完全明确，只是普遍认为有两大原因：一是吸烟，二是遗传因素．当前对于肺癌的诊断方法代价昂贵且误诊率较高；作为主要治疗手段的化疗，通常情况下不能完全治愈肺癌，只能延长患者生存周期和改善生活质量，往往还需要承受化疗所带来的副作用伤害．因此，肺癌风险致病基因的预测，不仅有助于全面彻底地了解疾病机理和遗传信息，还对疾病诊断、复发转移监测，以及疗效和预后判断，都有着十分重要的指导意义．

随着大量生物网络数据不断涌现，很多相关的计算研究也已展开，如生物网络比对、同源分析、路径挖掘等．相比于生化或临床实验方法，计算方法实验周期短、消耗人力物力少，因而利用计算方法预测风险致病基因已成为人类复杂疾病研究的热点和趋势．

重启游走(RWR)算法具有计算复杂度较低、空间消耗较小、所需先验信息较少以及结果较为全面等优点，因此进行全局风险致病基因预测时，最常用的是 RWR 算法以及在此基础上改进的各算法[4-11]．但这些方法普遍存在以下局限性：a．起始节点只简单考虑了已知致病基因，一方面相对于整个庞大网络而言致病基因节点数量可能太少，另一方面致病基因节点或整个网络的拓扑连接性可能较差，因此对于致病基因节点数量较少或拓扑连接较

* 国家自然科学基金(91430111, 61473232, 61170134)和国家自然科学基金青年基金(61502396)资助项目, 互联网金融创新及监管四川省协同创新中心资助项目.
** 通讯联系人.
Tel: 029-88431308, E-mail: zhangsw@nwpu.edu.cn
收稿日期：2015-12-07，接受日期：2016-01-14

差以及网络连接较为稀疏情况的游走效果不是很理想；b．转移矩阵大都是归一化的网络邻接矩阵或归一化的蛋白质相互作用强度矩阵，也就是说从起始节点游走到所有邻居节点的概率是等同的，这不仅完全依赖于网络拓扑性，而且没能反映出网络不同区域的拓扑特性，此外对于疾病的特有信息没有充分的予以利用；c．游走算法大都只简单地采用了游走结果，缺乏进一步生物学意义的验证，因此会对所得结果的显著性有一定的影响.

鉴于此，我们提出了一种新的基于扩展模糊度量以及拓扑相似度和差异相关性转移矩阵加权融合策略的风险致病基因预测算法(augmenting fuzzy measure and fusing similarity and correlation，AFMFSC)，并在肺癌疾病数据上验证算法有效性.

1　材料与方法

1.1　数据集

本文采用肺癌致病基因、人类蛋白质相互作用、表型相似性、基因表达谱及 GO (Gene Ontology) 5 个数据集．致病基因数据集由两部分组成，一部分来源于 LCD(lung cancer database)数据库，另一部分从 OMIM Gene Map 数据库筛选得到；人类蛋白质相互作用数据集来源于 HPRD 数据库(版本 9)，去除蛋白质自身相互作用，得到一个包含 9 502 个蛋白质，37 520 个相互作用的网络；表型相似性数据来自 Van Driel 等[12]的研究结果，以矩阵形式给出，包含了 5 080 种 OMIM 疾病表型以及表型之间的相似性得分；基因表达谱数据 GSE4115 和 GSE23066 来源于 GEO 数据库，其中 GSE4115 包含 79 个患有肺癌的吸烟者和 73 个无肺癌的吸烟者的样本，GSE23066 包含有肺癌遗传家族病史正常的 5 个样本和患病的 5 个样本.

1.2　AFMFSC 算法

AFMFSC 算法将与肺癌表型具有一定近似性的基因在 GO 功能注释层面上，借助增广模糊思想测量了基因间的功能相似性，并从中选出重要基因与原致病基因一同作为扩展的起始节点；然后利用 PPI 网络拓扑特性构建了拓扑相似度转移矩阵以及利用基因表达谱数据构建了基因表达差异相关性转移矩阵，分别用改进的转移矩阵和扩展的起始节点进行重启随机游走，并将两种游走结果进行加权融合；最后将融合结果进行排序并选取排名靠前的基因进行富集分析，最终确定出肺癌风险致病基因．AFMFSC 算法流程图如图 1 所示.

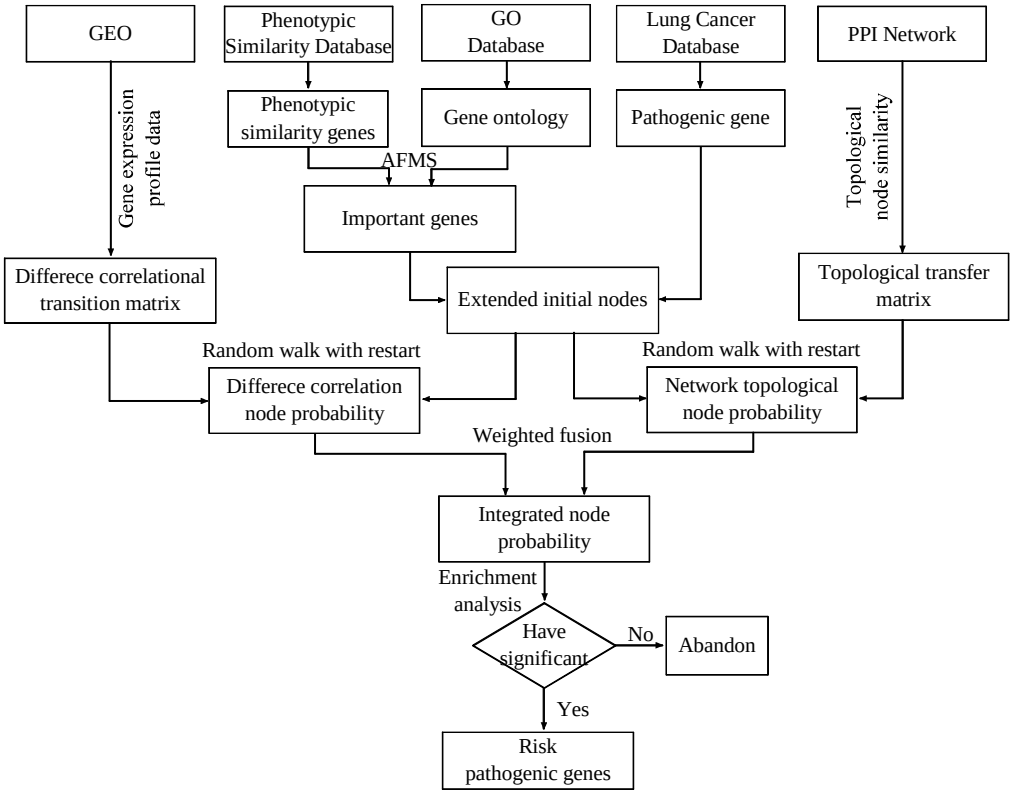

Fig. 1　The AFMFSC algorithm flow chart

1.2.1 基于增广模糊测量的扩展起始节点

随着研究的深入，发现没有共同注释术语的基因间也存在有一定的相似性．之前认为，如果两个基因没有共同的注释术语，则这两个基因间的相似性为 0，这一观点存在着很大的局限性[13-15]．因此，我们采用文献[16]的增广模糊测量相似性(augmenting fuzzy measure similarity，AFMS)思想，度量基因间的相似性．

在 GO 数据库中，一个基因有一个或多个术语对其进行注释．假设 G 代表某个基因，T_i 代表术语，G={T_1, ……, T_n}表示描述该基因的基本术语集合．定义 g 为模糊测量函数，它是一个具有真实值的函数．该函数具有以下性质：

①$g(\varnothing)=0$ 且 $g(G)=1$；

②假设 A, B 是两个不同的术语，如果 $A,B\subseteq G$ 且$A\subseteq B$，那么存在：$g(A)\leqslant g(B)$；

③对所有的 A, $B\subseteq G$ 且 $A\cap B=\Phi$ 则有：$g(A\cup B)=g(A)+g(B)+\lambda g(A)g(B)$．其中，当 $\lambda>-1$ 时成立．

定义模糊密度值 $g^i=g(\{T_i\})$，对 GO 中的每一个术语，它们对应的模糊密度值都可以采用文献[17]所提的方法计算得出．假设某个术语 T_i，在 GO 数据库中其自身及它的子术语出现的次数可以用概率 $p(T_i)$来表示：

$$p(T_i)=\frac{\text{数目}(T_i+T_i\text{ 的子术语})}{\text{数目}(\text{所有 GO 数据库中的术语})} \tag{1}$$

其中，$1\leqslant i\leqslant|GO|$.

模糊密度值 g^i 计算公式为：

$$g^i=-\ln(p(T_i))/\max_{T_j\in GO}\{-\ln(p(T_j)\} \tag{2}$$

显然，$g^i\in[0, 1]$.

在得到某个基因所有术语的模糊密度值后，性质③中的参数 λ 就可根据该性质以及 $G=\bigcup_{i=1}^{n}\{T_i\}$ 和 $g(G)=1$ 被唯一地确定出，计算公式如下：

$$1+\lambda=\prod_{i=1}^{n}(1+\lambda g^i) \tag{3}$$

从公式(3)中可以得到 λ 的唯一解，且能满足 $\lambda>-1$[16]．因此对于任一个在 GO 中有术语描述的基因 G_i，它对应的唯一参数 λ 就可以通过公式(3)计算出．

假设两个基因 G_1 和 G_2，在 GO 中它们的术语集分别为 $G_1=\{T_{11}, T_{12}, \cdots T_{1i}\}$ 和 $G_2=\{T_{21}, T_{22}, \cdots T_{2j}\}$，其中 $\forall T_{1i}$ 或 $T_{2j}\in$GO，它们术语的增广集分别定义为：

$$G_1^+=G_1\cup\{T_{1i,\,2j}\}\text{和 }G_2^+=G_2\cup\{T_{1i,\,2j}\} \tag{4}$$

其中，$\{T_{1i,\,2j}\}$表示术语对$(T_{1i,\,2j})$最近的公共父节点．扩展后 G_1^+或 G_2^+对应的参数 λ_1^+或 λ_2^+可通过公式(1)～(3)计算得到．进而可得两个基因的增广术语交集为：

$$[G_1\cap G_2]^+=G_1^+\cap G_2^+=[G_1\cap G_2]\cup\{T_{1i,\,2j}\} \tag{5}$$

利用增广术语交集，增广模糊测量相似性 $S_{\text{AFMS}}(G_1, G_2)$定义为：

$$S_{\text{AFMS}}(G_1, G_2)=\frac{g_1^+([G_1\cap G_2]^+)+g_2^+([G_1\cap G_2]^+)}{2} \tag{6}$$

其中，g_k^+是根据性质③以增广术语交集对应参数 λ_k^+和公式(5)扩展结果的增广模糊测量计算，$k=\{1, 2\}$.

对多个基因而言，可以通过 AFMS 方法计算出任意两个基因间的相似性得分，进而得到相似性得分矩阵 $SS(s_{ij})_{n\times n}$，其中 $s_{ij}=S_{\text{AFMS}}(G_i, G_j)$且当 $i=j$ 时，$s_{ij}=1$，n 表示基因数目．

基于 $SS(s_{ij})_{n\times n}$ 各基因的得分 s_i 定义为：

$$s_i=\frac{2(\sum_{j=1}^{n}s_{ij}-1)}{\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}-n} \tag{7}$$

可根据得分高低进行排序，选取所有基因得分总和中 90%的前约 m 个基因作为重要基因．

对肺癌疾病，可以借助疾病表型相似性和运用该方法，筛选重要基因和扩展起始节点．首先将肺癌表型所对应的表型 MIM 编号映射到表型相似性得分矩阵中，并将该表型与其他所有表型的相似性得分平均值作为阈值．挑选相似性得分大于此阈值的其他表型都将作为近似表型，并在 OMIM 中找到各相似表型所对应的基因；然后计算这些基因间的功能相似性，进而打分选出重要基因；最后同肺癌已知致病基因一同构成扩展起始节点集，成为 AFMFSC 算法的扩展起始节点集．并规定中的所有节点都拥有相同的概率重启，因此 AFMFSC 算法的初始向量 P^0 为：

$$P_{(v_i)}^0=\begin{cases}1/N(ES) & \text{如果 } v_i\in ES\\ 0 & \text{其他节点}\end{cases} \tag{8}$$

其中 $N(ES)$代表节点集 ES 的数目．

1.2.2 基于网络拓扑的转移矩阵

概率转移矩阵如果是简单的行或列归一化网络邻接矩阵，将导致同一个源节点跳转到其所有邻居节点的概率是一样的．但不同邻居节点在网络中的

特性是不一样的，因而并不合理．如邻居节点度的大小有所不同，因此对不同连接节点应该区分并给予不同的转移概率．AFMFSC 算法从网络拓扑特性出发对节点间的相似度进行测量，然后根据节点间的相似度值来重新构造节点到各邻居节点的跳转概率．

依据节点间的拓扑连接关系，Leicht 等[18]提出了一种网络拓扑节点相似度测量方法，其核心思想是：如果节点 i 的邻接节点 v 和 j 节点相似，那么则认为节点 i 也与节点 j 相似．首先假设网络中的每个节点都与自身是完全相似的，则节点 i 与节点j 的相似度值 S_{ij} 可表示为：

$$S_{ij}=\phi\sum_{v}A_{iv}S_{vj}+\psi\delta_{ij} \tag{9}$$

其中 A 是网络的邻接矩阵，参数 ϕ 和 ψ 是调节系数，δ_{ij} 是克罗内克函数，具体定义为：

$$\delta_{ij}=\begin{cases}0 & \text{如果 } i\neq j\\ 1 & \text{如果 } i=j\end{cases} \tag{10}$$

将公式(9)写成矩阵形式为：

$$S=\phi AS+\psi I \tag{11}$$

其中 I 是单位矩阵，S 为节点相似度矩阵．因为上式可以变为 $S=\psi[I-\phi A]^{-1}$，参数 ψ 仅仅是一个整体的倍数因子，相似性在本质上并不是一个绝对量级而只是不同节点对之间的相对量级，因此将该参数设为 1，因而进一步对上式简化为：

$$S=[I-\phi A]^{-1} \tag{12}$$

通过对公式(12)的幂级数展开以及利用复杂网络的平均路径长度理论，最终推出参数 $\phi=\varepsilon/\lambda_1$，其中 ε 是个常数且 $\varepsilon\in(0.9, 1)$，通常情况下一般取 ε=0.95，λ_1 表示邻接矩阵 A 的最大特征值．由于该算法仅是对转移矩阵的改进，所以只考虑目标节点及其邻接节点之间的相似度，对网络中没有边连接的节点对之间的拓扑相似度不予考虑．因此，进一步对邻接矩阵 A 和相似度矩阵 S 做 Hadamard 积运算，如公式(13)所示：

$$W^S=S\circ A \tag{13}$$

这里 W^S 表示改造过的转移矩阵，同时为了保证从节点跳转到所有邻接节点的总概率和为 1，还对 W^S 做了列归一化处理．

相应地，基于网络拓扑相似度的改进重启随机游走可用以下公式表示：

$$P_S^{t+1}=(1-\gamma)W^SP^t+\gamma P^0 \tag{14}$$

1.2.3　基于基因表达差异相关性的转移矩阵

疾病作为一种特殊的表达形式，其自身就包含有一定价值的特定信息．对于肺癌疾病，有包含 N 个基因的表达谱数据，则可用 $Z_i^d=[z_1^d, z_2^d, \cdots, z_l^d]$和 $Z_i^h=[z_1^h, z_2^h, \cdots, z_\tau^h](i\in N)$，分别代表疾病样本和正常样本的数目，如果在某些特殊情况下二者样本数目不相同，即 $l\neq\tau$，则需要对缺失的基因表达值做补 0 处理．相应地，基因表达值之差定义为：

$$\Delta Z_i=[(z_1^d-z_1^h), (z_2^d-z_2^h), \cdots, (z_m^d-z_m^h)] \tag{15}$$

其中：i=1, 2, ⋯, N，m=max(l, τ).

根据皮尔逊相关系数计算基因 i 与基因 j 各自的表达差值间的相关性，如公式(16)所示：

$$r_{ij}=\frac{Cov(\Delta Z_i, \Delta Z_j)}{\sigma_{\Delta Z_i}\sigma_{\Delta Z_j}} \tag{16}$$

其中，i, j=1, 2, ⋯, N，$\sigma_{\Delta Z_i}$和 $\sigma_{\Delta Z_j}$分别是 ΔZ_i 和 ΔZ_j 的标准差，Cov 是这二者的 ψ 协方差．进而可得到基因表达差异相关性矩阵 R.

由于依旧仅考虑网络中有连接的节点间的基因表达差异相关性，所以同样对相关性矩阵和网络的邻接矩阵 A 作 Hadamard 运算，如公式(17)所示：

$$W^c=R\circ A \tag{17}$$

为了使改造的转移矩阵满足一阶马尔科夫模型，依旧对矩阵 W^c 做列归一化处理，从而得到基于基因表达差异相关性的概率转移矩阵．对应的改进重启随机游走模型为：

$$P_c^{t+1}=(1-\gamma)W^cP^t+\gamma P^0 \tag{18}$$

1.2.4　加权融合

在游走过程中当时间趋于无穷且相邻的两次游走的网络节点概率的一范数小于 10^{-10} 时，则认为该游走趋于收敛，网络处于稳定状态．此时由公式(14)和公式(18)分别获得基于网络拓扑相似度和基因表达差异相关性的节点概率向量 P_s^∞ 和 P_c^∞．该概率向量是网络中节点与起始节点相似性度量的反映．为获得最终唯一结果，我们采用加权融合策略对 P_s^∞ 和 P_c^∞ 这两个概率向量进行融合，如公式(19)表示为：

$$P=\alpha P_s+(1-\alpha)P_c \tag{19}$$

P 表示融合后的节点综合概率向量，参数 $\alpha\in[0, 1]$是一个调节系数，用于调节网络拓扑相似度和基因表达相关性的平衡．

1.3　**算法步骤**

AFMFSC 算法步骤如下：

Step *1*，收集已知致病基因，在表型相似性数据中找出满足一定近似性要求的表型及其对应的

基因；

Step *2*，计算表型近似基因间的相似性，构建相似性得分矩阵和对各基因进行打分，从中选出重要基因并与已知致病基因一同构成扩展初始节点集；

Step *3*，在 PPI 中网络进行计算得到拓扑相似度矩阵，作为改进的拓扑转移矩阵；

Step *4*，对基因表达谱中差异相关性进行计算，结合 PPI 网络得到基因表达差异相关性矩阵，作为改进的差异相关性转移矩阵；

Step *5*，采用扩展起始节点集和改进的转移矩阵进行游走计算，分别获得基于网络拓扑相似度和基因表达相关性下的节点概率向量；

Step *6*，将两种不同的节点概率向量采用加权融合的策略进行融合，得到综合结果；

Step *7*，将结果按照大小进行排列，将排名前 Top 1%的基因进行富集分析，选择满足条件的基因作为最终的风险致病基因.

2 结果与讨论

2.1 模块挖掘结果及富集分析

在致病基因集中，我们共收集了 253 个肺癌致病基因. 其中 113 个致病基因来自于 LCD 数据库，140 个致病基因来自于 OMIM 数据库. 通过表型近似性筛选，得到 139 个近似基因，按照功能相似性测量及打分后，102 个重要的基因被选出，因而最终 AFMFSC 算法中扩展的起始节点集 *ES* 包含 355 个节点. 按照 AFMFSC 算法步骤在 PPI 网络中进行全局游走和打分排名后，我们选取排名前 Top 1%的基因在显著性阈值 *P*-value 为 0.01 的条件下进行 GO 和 KEGG 通路富集分析并选取满足条件的风险致病基因. 最终得到 73 个具有一定生物学意义的肺癌风险致病基因.

由于篇幅所限无法将预测出的所有风险致病基因进行功能和作用描述，因此只将排名前 10 的风险致病基因借助在线数据库和现有文献检索对其进行了描述. 它们大都有相关的研究文献予以支持，详细描述见表 S1 并且完整的风险致病基因名称见表 S2.

在富集分析过程中，借助于生物信息数据平台 DAVID[19]，对排名靠前的基因在生物学过程(BP)、细胞成分(CC)、分子功能(MF)和 KEGG 通路等方面进行了富集分析和确定. 我们将部分具有高富集生物学意义的结果 ($P < 0.01$)进行展示并进行简单分析，如表 1 和 2 所示：

Table 1 The partial results of GO enrichment analysis for lung cancer risk pathogenic genes

	GO ID	GO Term	Number of genes	Genome frequency	*P*-value
BP	GO:0010941	Regulation of cell death	67	13400/47198	2.83×10^{-46}
	GO:0043067	Regulation of programmed cell death	63	9695/47198	5.68×10^{-46}
	GO:0042981	Regulation of apoptosis	64	9750/47198	1.16×10^{-45}
	GO:0010033	Regulation of cell proliferation	55	7921/47198	1.72×10^{-41}
	GO:0007050	Cell cycle arrest	33	8206/47198	5.87×10^{-39}
	GO:0010604	Positive regulation of macromolecule metabolic process	48	7497/47198	8.21×10^{-38}
	GO:0031328	Positive regulation of cellular biosynthetic process	53	6730/47198	1.78×10^{-37}
	GO:0009719	Response to endogenous stimulus	42	6447/47198	4.79×10^{-35}
	GO:0016055	Wnt receptor signaling pathway	48	5271/47198	6.72×10^{-33}
CC	GO:0005615	Extracellular space	59	11383/47198	4.91×10^{-40}
	GO:0005911	Cell-cell junction	32	9992/47198	3.41×10^{-36}
	GO:0005829	Cytosol	21	17294/47198	2.77×10^{-26}
	GO:0009986	Cell surface	17	9417/47198	8.81×10^{-26}
	GO:0044459	Plasma membrane part	14	9061/47198	5.60×10^{-20}
MF	GO:0042802	Identical protein binding	53	9377/47198	2.66×10^{-24}
	GO:0019899	Enzyme binding	49	9964/47198	9.13×10^{-22}
	GO:0046983	Protein dimerization activity	47	9561/47198	1.96×10^{-19}
	GO:0004672	Protein kinase activity	34	8680/47198	3.31×10^{-18}
	GO:0043566	Structure-specific DNA binding	51	9288/47198	2.76×10^{-15}

Table 2　The partial results of KEGG enrichment analysis for lung cancer risk pathogenic genes

KEGG ID	KEGG Term	Number of genes	*P*-value
hsa05200	Pathways in cancer	72	1.63×10^{-45}
hsa05212	Pancreatic cancer	63	2.22×10^{-36}
hsa05222	Small cell lung cancer	67	1.22×10^{-33}
hsa04012	ErbB signaling pathway	62	1.15×10^{-29}
hsa05220	Chronic myeloid leukemia	57	5.82×10^{-27}
hsa05210	Colorectal cancer	50	8.19×10^{-23}
hsa05223	Non-small cell lung cancer	58	7.54×10^{-21}
hsa05219	Bladder cancer	43	8.18×10^{-19}

从表 1 和表 2 中可以看出，这些风险疾病基因在 GO 富集分析中大都表现出跟细胞活动和细胞器的死亡、周期、细胞内信号、细胞核、蛋白质等相关，在 KEGG 路径富集分析中都表现出与多种癌症疾病有密切的关系．这是由于癌症细胞本质上来源于组织和机体的正常细胞，只是在细胞有丝分裂的不同阶段受到各种因素或信号的影响，导致其出现分裂和增殖异常．癌症最明显的标志就是癌细胞生长失控，同时癌细胞还相互诱发和转移．这充分说明这些预测结果对肺癌疾病具有较强的现实生物学意义和可信度．例如：GO:0043566 与肺癌基因在关键区域 3p21.3 的聚集有关，GO:0005911 与肺细胞肺癌的钙黏蛋白和连环蛋白表达有关，GO:0007050 与小细胞肺癌中 Notch 信号诱发细胞周期停滞有关，文献[20]的研究表明 Wnt 信号路径对 A549 细胞中的肺癌干细胞起着重要的上调控作用．路径 hsa05222 和 hsa05223 直接涉及到了小细胞和非小细胞肺癌疾病中，其他癌症路径中的基因也与肺癌有着较为密切的联系，例如肺癌中非正常的消亡和增殖等等．

2.2　参数对算法性能影响

AFMFSC 算法中涉及 γ 和 α 两个参数．其中参数 $\gamma(0\leqslant\gamma\leqslant1)$用于调节游走过程中疾病的先验信息(已知致病基因和所扩展的重要基因)和网络拓扑的重要性．参数 γ 越大则表示每次迭代过程中重启的概率就越大，相应的游走对于疾病的先验信息的依赖就越强，而对于网络拓扑特性的依赖就弱．参数 $\alpha(0\leqslant\alpha\leqslant1)$用于对基于网络拓扑相似度游走得到的节点概率和基于基因表达谱数据相关游走得到的节点概率进行加权融合．当 α=0 时 AFMFSC 算法的游走过程只受到基因表达谱相关性的引导，当 α=1 时节点的概率仅由网络拓扑特性的游走决定．在这里，我们通过留一验证法绘制 ROC 曲线，按照 AUC 指标分别来衡量这 2 个参数对 AFMFSC 算法的性能影响．

参数 γ 以 0.1 大小为间隔从 0.1 变化到 1 的过程中，可得 AFMFSC 算法下不同的 AUC 值，其关系曲线如图 2 所示．需要说明的是，为了确保公正性，在计算过程中我们将参数暂设为 0.5．从曲线可以看出，参数 γ 取值在 0.1～0.6 之间随着它的增长 AUC 值呈上升趋势；但在取值 0.6～1 之间随着它的继续增长时，AUC 值却逐步下降；但是 γ 的取值在 0.4～0.8 之间时，对应变化的差异并不十分显著．我们选择 AUC 值最大时所对定的参数值，因此在肺癌疾病数据条件下公式(14)和公式(18)的参数 γ 最佳选择为 0.6.

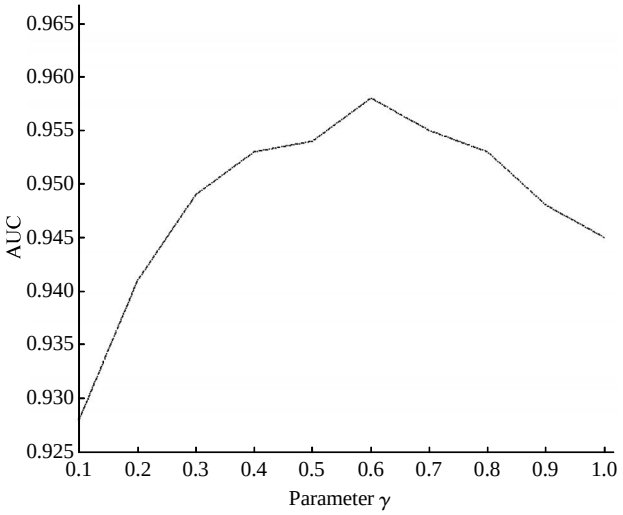

Fig. 2　The relationship between parameter γ and AUC of AFMFSC algorithm

此外，我们还考察了该参数对 RWR 算法 AUC 指标的影响，其关系曲线如图 3 所示．从曲线也可以看出，AUC 值同样在参数在 0.1～0.6 的范围内逐步上升，在 0.6～1 的范围内逐步下降．综合这两种情况，我们可以得出一个结论，参数 γ 对重启游走类的算法的总体影响趋势是一致的，即不管 RWR 算法还是 AFMFSC 算法，参数 γ 对其 ROC 曲线的 AUC 指标影响趋势是一致的，都有其性能先上升后下降，存在一个性能峰值的情况．虽然在不同的算法下具体的结果和性能可能有所不同，但总体趋势的差别不大．

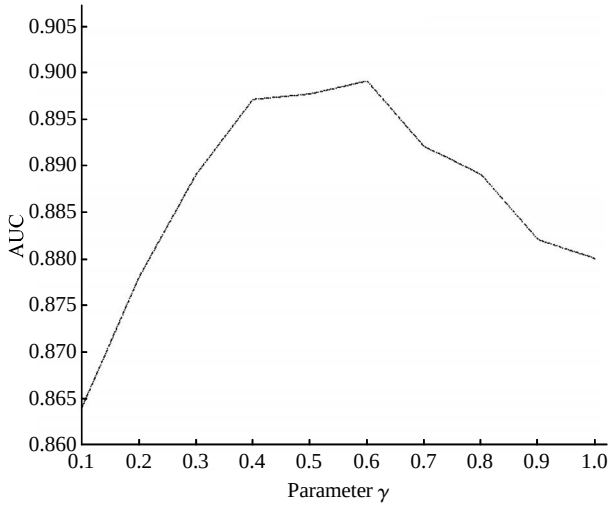

Fig. 3　The relationship between parameter γ and AUC of RWR algorithm

同样通过设置不同的 α 值($\alpha \in (0, 1)$)来考察其对 AFMFSC 算法 AUC 指标的影响，此时取参数 γ=0.6，其二者关系曲线如图 4 所示．从图中可以看出，参数 α 对 AUC 指标的结果有着较大的影响．当 $\alpha \in [0, 0.69]$时，AFMFSC 算法的 AUC 值呈上升趋势；当 $\alpha \in [0.69, 1]$时，AUC 值呈下降趋势且下降较为迅速．此外我们还观察到当 $\alpha \in [0.5, 0.8]$时，AUC 值虽有一定的波动，但整体波动的幅度有限，总体可以说是较为稳定的，且在 α=0.69 时 AUC值达到最大，说明此时算法性能最优．因此在肺癌疾病数据条件下公式(19)中参数取为 0.69．

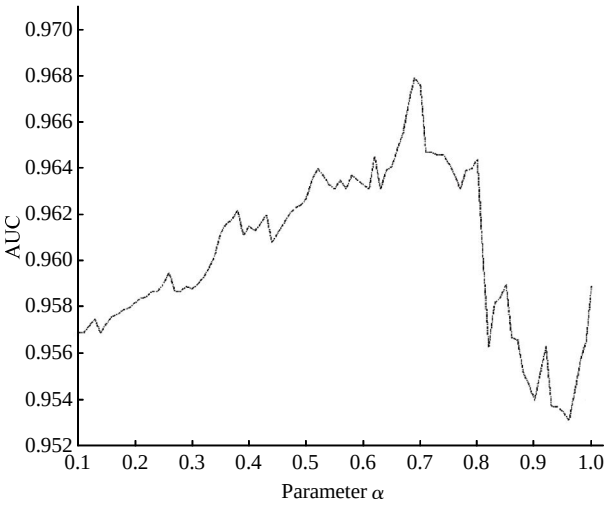

Fig. 4　The relationship between parameter α and AUC of AFMFSC algorithm

2.3　与其他算法性能比较

AFMFSC 算法最佳参数确定后，我们用留一验证法绘制出 ROC 曲线．为了证明该算法的有效性，还将其与 RWR 算法[4]、PRINCE[5]、ORIENT[6]算法以及张等[11]所提的 ARWRH 算法进行比较，也绘制其相应的 ROC 曲线，如图 5 所示．除 AUC

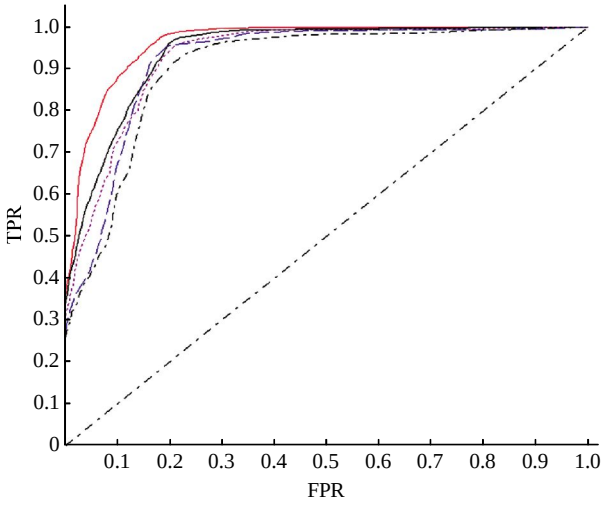

Fig. 5　The ROC curves of AFMFSC, ARWRH, ORIENT, PRINCE, and RWR algorithm

—: AFMFSC(AUC≈96.22%); —: ARWRH(AUC≈94.03%); ----: ORIFNT(AUC≈92.98%); —: PRINCE(AUC≈91.23%); – –: RWR (AUC≈90.33%).

值之外，还考虑了 Top 1%，Top 5%和平均排名等这些指标，分别用 R_P^1，R_P^5和 R_v 表示．Top 1%表示用留一验证法的排序测试基因中排名小于或等于测试基因数目的 1%的基因中真实阳性数据所占的比例，Top 5%的表示方法也类似如此，这两个性能指标越大，则说明算法预测精度越高；平均排名表示所有真实阳性数据排名的平均值，平均排名越小则算法性能越优秀．表 3 列出了不同游走算法下所对应的 AUC，R_P^1，R_P^5和 R_v 等指标值．

Table 3　The performance of AFMFSC, ARWRH, ORIENT, PRINCE, and RWR algorithms

Method	AUC	R_P^1	R_P^5	R_v
RWR	0.903	28.13	61.67	9.59
PRINCE	0.912	29.92	59.88	9.28
ORIENT	0.930	33.71	64.52	8.84
ARWRH	0.940	34.64	67.36	6.27
AFMFSC	0.962	36.23	70.15	5.23

从图 5 和表 3 中可以看出，AFMFSC 算法相对于其他 4 种算法有着较高的敏感性和特异性，AUC 值比其他 4 种算法分别提高了 0.0219、0.0324、0.0499 和 0.0589，在 Top 1% 和 Top 5%性能指标上 AFMFSC 算法表现出明显的优势，此外算法的平均排名也分别比其他 4 种算法提高了 1.04、3.61、4.05、4.36．这些数据充分表明 AFMFSC 算法在风险致病基因预测的精确度方面比其他 4 种游走算法更为优异．

2.4　转移矩阵对 AFMFSC 算法性能影响

为了分析单一转移矩阵重启游走算法(邻接矩阵 A、拓扑相似度矩阵 W^s 或表达差异相关性矩阵 W^c)与融合多转移矩阵结果的 AFMFSC 算法间的性能差别，我们同样使用留一验证法，分别考察了不同转移矩阵和相同扩展起始节点下游走算法的 AUC、Top 1%、Top 5%和平均排名等指标．表 4 列出了不同转移矩阵下所对应的 AUC、R_P^1、R_P^5和 R_v．从表中可以看出，3 种单一转移矩阵游走算法的 AUC、R_P^1、R_P^5和 R_v 相差不明显，甚至单一的以 W^s 或 W^c 为转移矩阵比直接以邻接矩阵作为转移矩阵的游走效果在某些方面还略差，但融合后整体性能却得到了明显的提高，另外还发现 RWR 算法在扩展节点的条件下，AUC 有了明显提高．这些结果说明：一方面基于网络拓扑相似度游走和基于基因表达谱差异相关性游走对网络节点间转移概率的弱依赖性；另一方面说明 AFMFSC 算法中对起始节点的扩展改进在风险致病基因预测的精度上有较大的提升．这种现象从对立的角度表明基于网络的方法对网络节点间更深层次的相似有着较强的依赖性．此外，加权融合对算法整体的性能有明显的提升，这也说明 W^s 或 W^c 对 AFMFSC 算法游走过程中信息流的引导是互补的．

Table 4　The performance of algorithm using extended initial nodes and different transfer matrix

Matrix	AUC	R_P^1	R_P^5	R_v
A	0.927	32.16	63.23	7.67
W^s	0.932	32.75	64.06	7.28
W^c	0.914	29.34	62.31	8.13
W^s+W^c	0.962	36.23	70.15	5.11

2.5　节点度偏差对 AFMFSC 算法性能影响

研究表明，基于信息流的方法对预测目标的节点度都有着不同程度的偏差[21]．也就是说，这些方法在进行风险致病基因预测时易受网络拓扑特征影响，对节点度较低的基因普遍存在精度较低，偏差较为明显的现象．AFMFSC 算法在扩展起始节点时，使用 GO 功能相似性而非网络拓扑性，并且在转移矩阵中融入了基因表达谱等额外的信息，因而这些策略都能在一定程度上弥补这种偏差．为了验证不同算法对节点度的影响，我们运用 AFMFSC、ARWRH、ORIENT、PRINCE 及 RWR 算法，以扩展起始节点集 ES 为对象，用留一法和相应的排序测试基因进行了验证，分别计算了不同节点度的基因的平均排名大小，并绘制了二者间的关系图，如图 6 所示．从图中可以看出，这些算法在不同节点度下的平均排名整体趋势都较为近似，但 AFMFSC 算法在各节点度条件下的平均排名相比于其他算法普遍较小，尤其是节点度在 1～4 范围内的平均排名明显最小．这说明：AFMFSC 算法在预测风险致病基因时受网络节点拓扑特征偏差的影响最小，所得结果的精确度和可信度更高．

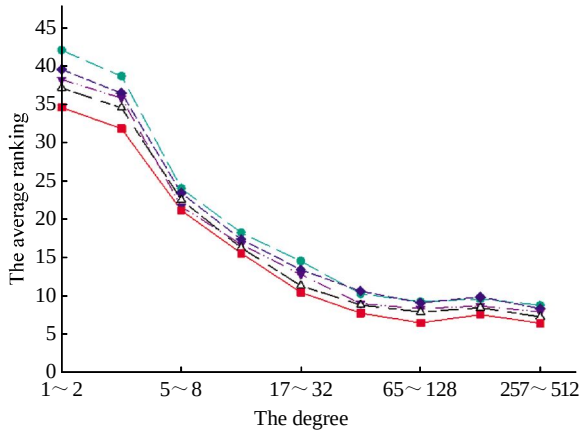

Fig. 6　The relationship between the degree and the average ranking of genes for different algorithms

■—■ : AFMFSC; △--△ : ARWRH; ▼----▼ : ORIFNT; ◆----◆: PRINCE; ●----● : RWR.

3　结　　论

针对重启随机游走框架预测风险致病基因时普遍存在起始节点较少、节点转移概率相同、信息源单一的问题，提出一种基于扩展起始节点和加权融合策略的改进重启随机游走风险致病基因预测算法(AFMFSC). 该算法利用增广模糊测量思想对疾病表型近似的基因进行打分，从中选出重要基因作为可扩展节点，构建了拓扑相似度矩阵和表达差异相关性矩阵作为新的转移矩阵并分别进行了重启游走，最后将两种结果进行加权融合和富集分析来确认风险致病基因. 肺癌实验结果表明 AFMFSC 算法所挖掘出的肺癌风险致病基因与肺癌有着密切联系，在疾病发生、发展过程中起到了一定的作用和影响，有着较好的生物学意义. 与其他游走算法相比，AFMFSC 算法的 Top 1%、Top 5%和平均排名等指标较高，AUC 值也大于其他方法；算法融合策略的性能也明显优于单一转移矩阵游走性能；AFMFSC 算法中各节点度的平均排名整体小于其他算法，尤其是节点度在 1～4 范围内的平均排名具有明显优势，说明 AFMFSC 算法在风险疾病基因预测时受网络拓扑特征偏差影响程度较小，所得结果准确可靠.

综上所述，AFMFSC 算法对于风险致病基因预测而言，算法简单易实现、结果有效可靠、不仅能预测出具有一定生物学意义的肺癌风险致病基因，有助于该疾病的诊断、预防和治疗，还可扩展预测其他疾病风险致病基因.

附件　表 S1，S2 见本文网络版附录(http://www.pibb.ac.cn).

参 考 文 献

[1] DuPage M, Jacks T. Genetically engineered mouse models of cancer reveal new insights about the antitumor immune response. Current Opinion in Immunology, 2013, **25**(2): 192−199

[2] Siegel R, Ma J, Zou Z, *et al*. Cancer statistics, 2014. CA: a cancer journal for clinicians, 2014, **64**(1): 9−29

[3] e Zahra S N, Khattak N A, Mir A. Comparative modeling and docking studies of p16ink4/Cyclin D1/Rb pathway genes in lung cancer revealed functionally interactive residue of RB1 and its functional partner E2F1. Theoretical Biology and Medical Modelling, 2013, **10**(1): 1

[4] Köhler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics, 2008, **82**(4): 949−958

[5] Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease *via* network propagation. PLoS Comput Biol, **6**(1): e1000641(DOI: 10.1371/journal.pcbi.1000641)

[6] Le D H, Kwon Y K. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. Computational Biology and Chemistry, 2013, **44**(01): 1−8

[7] Guo X, Gao L, Wei C, *et al*. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. PloS one, 2011, **6**(9): e34171.

[8] Xie M, Hwang T, Kuang R. Reconstructing disease phenome-genome association by bi-random walk. Bioinformatics, 2012, **1**(02): 1−8

[9] Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics, 2010, **26**(8): 1057−1063

[10] Li Y, Patra J C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 2010, **26**(9): 1219−1224

[11] 张松瑶, 张绍武. 基于二元网络异步重启随机游走算法预测肺癌风险致病基因. 生物物理学报, 2015, **31**(1): 33−44
Zhang S Y, Zhang S W. Acta Biophysica Sinica, 2015, **31** (1): 33−44

[12] Van Driel M A, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. European Journal of Human Genetics, 2006, **14**(5): 535−542

[13] Lord P W, Stevens R D, Brass A, *et al*. Semantic similarity measures as tools for exploring the gene ontology//Pacific Symposium on Biocomputing. 2003, **8**(04): 601−612

[14] Speer N, Spieth C, Zell A. A memetic clustering algorithm for the functional partition of genes based on the gene ontology// Computational Intelligence in Bioinformatics and Computational

Biology, 2004. Proceedings of the 2004 IEEE Symposium. LosAlamitors, CA: IEEE Computer Society Press, 2004: 252−259

[15] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res (JAIR), 1999, **11**(05): 95−130

[16] Popescu M, Keller J M, Mitchell J A. Fuzzy measures on the gene ontology for gene product similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2006, **3** (3): 263−274

[17] Grabisch M, Sugeno M, Murofushi T. Fuzzy measures and integrals: theory and applications. New Yor: Springer-Verlag, 2000

[18] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks. Physical Review E, 2006, **73**(2): 026120

[19] Alvord G, Roayaei J, Stephens R, *et al*. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol, 2007, **8**(9): 183

[20] Zakaria N, Yusoff N M, Zakaria Z, *et al*. Human non-small cell lung cancer expresses putative cancer stem cell markers and exhibits the transcriptomic profile of multipotent cells. BMC Cancer, 2015, **15**(1): 84

[21] Erten S, Koyutü rk M. Role of centrality in network-based prioritization of disease genes//Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Heidelberg，Berlin: Springer-Verlag Press, 2010: 13−25

Uncovering Lung Cancer Risk Pathogenic Genes With Expanded Initial Node and Weighted Fusion Strategy*

WANG Yi-Bin, CHENG Yong-Mei, ZHANG Shao-Wu**

(*School of Automation, Key Laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical University, Xi'an* 710072, China)

Abstract The identification of risk pathogenic genes for lung cancer is helpful to understand disease pathogenesis and improve clinical practice. However, the present predicting methods of using RWR framework include the common problems of the less initial nodes, the same node transition probability, and the single information source. To further improve the performance of RWR framework, we propose a novel method named AFMFSC to identify disease-related genes, by enlarging the initial nodes and weighted fusion strategy, and use lung cancer as the test object. The AFMFSC algorithm first computes the augmented functional similarity scores between disease phenotype approximate genes based on the idea of augmenting fuzzy measure similarity, screens important genes as the expanded initial nodes together with pathogenic genes, then walks in the global PPI network separately guided by the node similarity transition matrix constructed with PPI network topological similarity properties and the correlational transition matrix constructed with the gene expression profiles, all the genes in the network are ranked by weighted fusing the above results guided by two types of transition matrices, at last the top ranked genes in the enrichment analysis as final risk pathogenic genes are determined. 73 significant genes are predicted to be the risk pathogenic genes for lung cancer, which are closely linked with the generation and development of this disease. Compared with the existing methods for prioritizing potential risk disease genes, the AFMFSC achieves a smaller average rank and less affect by degree distribution bias but bigger Top 1%，Top 5% and AUC value. In addition, the ranking performance of fusion strategy outperforms a single transfer matrix or ordinary adjacency matrix. The AFMFSC algorithm not only can accurately and effectively predict the risk pathogenic genes of lung cancer, but also can be easily extended to identify any other diseases related genes, and provide additional insights for exploring the pathogenesis of cancer.

Key words risk pathogenic gene, expanded initial node, topological similarity transition matrix, gene expression difference correlational transition matrix, random walk with restart

DOI: 10.16476/j.pibb.2015.0380

*This work was supported by grants from The National Natural Science Foundation of China (91430111, 61473232, 61170134), The National Natural Science Foundation of China Youth Fund Project (61502396), and the Internet Financial Innovation and Supervision of Collaborative Innovation Center in Sichuan Province.

**Corresponding author.

Tel: 86-29-88431308, E-mail: zhangsw@nwpu.edu.cn

Received: December 7, 2015 Accepted: January 14, 2016

附　录

Table S1　The function description of the top 10 risk pathogenic genes

Gene name	Gene description
IGF1R	The insulin-like growth factor I receptor plays a critical role in transformation events. It is highly overexpressed in most malignant tissues where it functions as an anti-apoptotic agent by enhancing cell survival.?Literature [Appendix 1] believes that IGF-1R expression was a negative predictive factor for a response to EGFR-TKIs in NSCLC patients harboring activating EGFR mutations.
PTEN	This gene was identified as a tumor suppressor that is mutated in a large number of cancers at high frequency. Literature [Appendix 2] has reported that decreased expression of PTEN in lung cancer PC9 cells harboring an EGFR-activating mutation results in acquisition of resistance to EGFR-TKIs.
TYMS	Expression of this gene and that of a naturally occurring antisense transcript rTSalpha vary inversely when cell-growth progresses from late-log to plateau phase. Literature[Appendix 3] believes that the gene expression level is shown to be an independent predictive factor in stage I and II NSCLC patients
SLC34A2	The protein encoded by this gene is a pH-sensitive sodium-dependent phosphate transporter. Defects in this gene are a cause of pulmonary alveolar microlithiasis. Literature [Appendix 4] believes that maintaining the reduced expression of SLC34A2/NaPi-IIb should provide benefits to LC cells.
RET	This gene encodes one of the receptor tyrosine kinases. This gene plays a crucial role in neural crest development, and it can undergo oncogenic activation in vivo and in vitro by cytogenetic rearrangement. Literature [Appendix 5] discoveries that there are significant changes in non-small cell lung cancer, and significant results can be obtained in patients with non-small cell lung cancer who carry KIF5B-RET recombination in a prospective clinical trial.
FOXP3	The protein encoded by this gene is a member of the forkhead family of transcriptional regulators. Defects in this gene are the cause of immunodeficiency polyendocrinopathy. Literature [Appendix 6] suggests that tumor FOXP3 expression has a better prognostic potential in NSCLC.
MAPK1	This gene encodes a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases, act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. ?This protein also acts as a transcriptional repressor independent of its kinase activity.
FGFR1	The protein encoded by this gene is a member of the fibroblast growth factor receptor family. Mutations in this gene are associated with a variety of diseases. Chromosomal aberrations involving this gene are associated with stem cell myeloproliferative disorder and stem cell leukemia lymphoma syndrome. Literature [Appendix 7] In vitro and in vivo studies have shown FGFR1-mediated signaling plays an important role in NSCLC cell growth, survival and migration, and shows “FGFR1-positive” status is an independent favourable prognostic factor in non-small cell lung cancer patients.
DACH1	This gene encodes a chromatin-associated protein that associates with other DNA-binding transcription factors to regulate gene expression and cell fate determination during development. Expression of this gene is lost in some forms of metastatic cancer, and is correlated with poor prognosis. Multiple transcript variants encoding different isoforms have been found for this gene.
MMP2	This gene is a member of the matrix metalloproteinase gene family. Activation of this protein can occur on the cell membrane. This protein is thought to be involved in multiple pathways including roles in the nervous system, regulation of vascularization, and metastasis.

Table S2　The list of risk pathogenic genes predicted by AFMFSC algorithm

Gene name
ACE ACHE ADAM9 BMP1 BTC CALR CCL2 CD40LG CHI3L1 CMTM7 COL1A1 DACH1 EDN1 EFEMP1 EGFR EREG FASLG FGF2 FGFR1 FLT1 FOXP3 GHR GHRL GPC5 HDGF HMOX1 HP HTRA1 IBSP ICAM1 IGF1R ISG15 KIT KITLG KLK8 LGALS1 LOXL1 MAPK1 MBL2 MICA MIF MMP2 MSR1 MUC16 NGF NID2 NRG1 NTN1 NUCB2 PECAM1 PGF PLA2G2A POSTN PRDX4 PROM1 PTEN PTHLH PVR RET SAA1 SFTPA1 SFTPC SLIT2 SPARC TDGF1 THPO TIMP1 TYMS VASH1 VEGFC VWF WFDC2 WNT1

参 考 文 献

[1] Yeo C D, Park K H, Park C K, *et al*. Expression of insulin-like growth factor 1 receptor (IGF-1R) predicts poor responses to epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors in non-small cell lung cancer patients harboring activating EGFR mutations. Lung Cancer, 2015

[2] Maeda M, Murakami Y, Watari K, *et al*. CpG hypermethylation contributes to decreased expression of PTEN during acquired resistance to gefitinib in human lung cancer cell lines. Lung Cancer, 2015

[3] Eguchi K, Oyama T, Tajima A, *et al*. Intratumoral gene expression of 5-fluorouracil pharmacokinetics-related enzymes in stage I and II non-small cell lung cancer patients treated with uracil-tegafur after surgery: A prospective multi-institutional study in Japan. Lung Cancer, 2015, **87**(1): 53−58

[4] Cerri M F, d Rezende L C, Paes M F, *et al*. Evaluation of relative expression of SLC34A2/NaPi-IIb in lung cancer cell lines treated with estrogen and PKC and PKA pathway modulators. Cancer Research, 2014, **74**(19 Supplement): 463−463

[5] Lipson D, Capelletti M, Yelensky R, *et al*. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nature medicine, 2012, **18**(3): 382−384

[6] Tao H, Mimura Y, Aoe K, *et al*. Prognostic potential of FOXP3 expression in non-small cell lung cancer cells combined with tumor-infiltrating regulatory T cells. Lung Cancer, 2012, **75** (1): 95−101

[7] Tran T N, Selinger C I, Kohonen-Corish M R J, *et al*. Fibroblast growth factor receptor 1 (FGFR1) copy number is an independent prognostic factor in non-small cell lung cancer. Lung Cancer, 2013, **81**(3): 462−467

[8] Govindan R, Ding L, Griffith M, *et al*. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell, 2012, **150**(6): 1121−1134