

一种基于 MeSH 主题词的临床-组学关系挖掘方法*

冯亚宁 焦梦映 段会龙 邓 宁**

(浙江大学生物医学工程与仪器科学学院, 生物医学工程教育部重点实验室, 杭州 310027)

摘要 癌症的发生发展与机体内基因的改变有密切联系, 在临床上表现为症状或检测指标的异常. 通过挖掘分析临床表现与基因改变之间的关系, 可为癌症早期诊断和精准治疗提供临床决策支持. 从文献数据出发, 利用结论性数据挖掘基因与临床表现的关系具有重要意义. 本文提出一种基于医学主题词(Medical Subject Headings, MeSH)的生物医学实体关系挖掘方法. 该方法利用 PubMed 中提供的文献信息, 借用向量空间模型思想, 使用 MeSH 主题词矢量表达待研究实体, 引入文献相互引用因素对结果进行修正, 将关系挖掘转化为矢量间的数学运算, 实现定量分析. 本文将该方法应用于结直肠癌临床表现和基因关系的研究中, 得到与结直肠癌相关的 203 个基因和对应的临床-基因 462 个关系. 通过结合使用基因功能和通路分析工具 g:Profiler 和 KEGG 等, 对结果进行分析验证. 结果表明, 基于 MeSH 主题词的文献挖掘方法, 避免传统“共现”方法对发现潜在关系的限制和复杂语义分析带来的大量计算, 为生物实体之间潜在关系的挖掘提供一种新的思路和方法.

关键词 文献挖掘, 定量分析, 关系挖掘, 结直肠癌, PubMed, MeSH

学科分类号 Q-3, R44

DOI: 10.16476/j.pibb.2015.0129

癌症目前已成为危害人类生命健康的主要疾病之一. 据美国癌症协会报道, 2015 年美国将新增癌症病患 1 658 370 例, 其中预计将有 589 430 病例死于癌症^[1]. 在世界卫生组织发表的《全球癌症报告 2015》中, 中国新增和死亡病例居世界之首. 癌症的发病机理被认为是多基因变异的联合效应, 包括原癌基因、抑癌基因的激活和灭活, 以及凋亡调节基因或 DNA 修复基因的改变^[2]. 基因的变异在临床上体现为正常组织的病变, 患者出现疼痛、溃疡、出血等症状, 以及多项检测指标的异常. 研究临床表现与基因变异之间的关系, 对癌症分子机制的认识, 获得正确及时的治疗方法, 具有重要意义.

结直肠癌(colorectal cancer, CRC)有相对于其他癌症更清晰的分子机制^[3], 利于对研究结果的检验, 成为研究癌症机制的对象之一. 目前, 大量结直肠癌相关研究结果以指数形式增长, 为肠癌研究提供了丰富的数据资源. 截止到目前为止, 生物医学文献库 MEDLINE^[4]已拥有超过 24 000 000 的文章标题和摘要信息, 包含大量丰富的生物医学研究成果. 如何运用这些文献进行临床-组学关系的二

次发现, 挖掘生物医学实体间(基因、蛋白质、临床术语、通路等)的关系, 成为生物医学信息学研究的关注点之一. 目前文献挖掘方法主要从两个方向开展, 一方面是基于“共现”原则的分析, 另一方面是基于统计学模型的挖掘, 如贝叶斯模型、向量空间模型等. 例如, Arzucan Özgür^[5]等用“共现”原则构建了一个基因-疾病网络. 他们将与前列腺癌相关的 10 个基因作为种子, 通过对 PubMed Central 中近 5 万篇文章全文进行分析, 找出与种子基因出现在同一篇文章的同一句话中的其他基因, 通过一系列数学指标的运算, 最后得到 EREG 等与前列腺癌相关的基因. Vijay Narayanasamy^[6]等提出的一个生物医学实体间关系挖掘系统 TransMiner, 利用“共现”原则筛选出同时出现基因 A 和疾病 B 的文章, 并利用 TF-IDF (term

* 国家自然科学基金(31100592), 国家高技术研究发展计划(863)(2012AA02A601)和国家科技重大专项(2013ZX03005012)资助项目.

** 通讯联系人.

Tel: 0571-87951792, E-mail: zju.dengning@gmail.com

收稿日期: 2015-04-29, 接受日期: 2015-07-15

frequency-inverse document frequency) 权值衡量 A 和 B 与文章的相关度, 结合数学运算, 挖掘出基因 - 疾病之间的关系。但目前的研究仍存在不足。首先, 基于“共现”原则的挖掘, 准确率依赖于大量语法和语义分析, 计算量大, 分析文章的数量范围有限; 第二, “共现”原则基于同时出现的条件, 着重于直接关系的分析, 限制潜在关系的发现。

本文提出了一种基于医学主题词 (medical subject headings, MeSH) 的文献挖掘方法。该方法借用“向量空间模型”的思想^[7], 将 PubMed 中提供的 MeSH 主题词作为特征项, 矢量表示待研究实体 (临床术语和基因), 并引入文献相互引用信息矢量表达结果, 最后结合矢量间的余弦相似性计算, 定量分析实体之间的关系。另外, 将关系挖掘有效转化为矢量间的数学运算, 在减少计算量的同时, 避免“共现”原则带来的限制, 对潜在关系的挖掘提供一种新的思路和方法。本文用该方法挖掘并分析与结直肠癌相关的 66 个临床术语与基因之间的关系, 通过接收者操作特征曲线 (receiver operating characteristic curve, ROC 曲线)^[8]对方法进行评估, 选取最佳工作点作为筛选阈值, 得到与 CRC 相关基因 203 个以及对应的 462 个关系, 并将结果与 GeneCards 中的文献挖掘方法——Novoseek 评分算法结果对比分析。最后使用工具 g:Profiler, KEGG 对得到的基因进行功能和通路的生物学验证。

1 方 法

1.1 数据获取

本文选取 MEDLINE 中与肠道疾病相关的的所有文章组成文献集, 利用 PubMed 提供的检索接口 E-Utilities^[9], 通过限定主题词为“Intestinal Diseases”, 下载获取结直肠癌相关文章信息, 以 xml 文件形式存放于本地。符合条件的文章共 473 242 篇 (截至 2014 年 6 月 30 日), 从中提取文章信息包括: 文章标题、摘要、发表年份、被引用文章 ID、MeSH 主题词等, 结构化处理后存放于本地数据库中。

应用于本文的所有结直肠癌相关临床表现, 统一表述为临床术语, 通过两个途径筛选获得: a. 通过结直肠癌患者的电子病历 (electronic health record, EHR) 提取, 包括临床症状、临床检测指标

等; b. 来自课题组前期工作中^[10]通过语义筛选的方法, 从高通量基因表达数据库 (gene expression omnibus, GEO) 中获取。其中包括了“Adenocarcinoma”、“High-Frequency Microsatellite Instability”等重要临床术语。最终获得 66 个与结直肠癌相关临床术语。附件表 S1 列出了本文中使用的临床术语。

本文采用的基因数据是从国际人类基因命名数据库 (HUGO Gene Nomenclature Committee, HGNC) 中获取的, 经认证的共 39 341 个记录, 下载存放于本地数据库^[11]。其中包括基因的认证符号 (Approved Symbol)、认证名称 (Approved Name)、曾用符号 (Previous Symbols)、曾用名称 (Previous Name) 等信息。

本文使用 E-Utilities 工具获取与临床术语或基因相关的文章信息。为保证准确性, 相关文章数量小于 3 篇的实体对象, 被排除在研究之外。最终筛选出 4 314 个基因和 45 个临床术语符合条件, 作为下一步研究对象。

1.2 方法设计

1.2.1 方法基本原理

本文借鉴了信息研究领域“向量空间模型”思想^[7], 用矢量表示待研究实体。向量空间模型 (VSM) 的基本思想是, 利用向量表示待研究文档, 即: $(w_1, w_2, w_3 \cdots w_n)$, 其中 w_i 表示第 i 个特征项的权重。文本分类中, 一般选取词或词组作为特征项, 权值通常用它们在文章中对应的频率表示。实际研究中, 相对词频较绝对词频能获得更精确的结果, 其中以 TF-IDF 算法^[12]最为经典, 其原理大致如下:

TF-IDF = 词频 (TF) × 逆文档频率 (IDF)

$$= \frac{\text{某词在文档中出现次数}}{\text{文档的总词数}} \times \lg \left(\frac{\text{语料库中文档总数}}{\text{包含该词的文档数} + 1} \right)$$

TF-IDF 算法思想主要从两个因素考虑, 一是特征项与文档相关的程度, 二是特征项在语料库中的表现力 (如“是”, “的”等属于表现力较弱的特征项, 而“结直肠癌”, “生物医学”等属于表现力较强的特征项)。向量空间模型是一个用来表示文本文件的代数模型, 现已应用于信息过滤、检索、分类、关联规则等领域中。特征项的提取通常需要去除表现力较弱的词, 提高计算的精确度。本文选取 PubMed 中提供的 MeSH 主题词为特征项集合。MeSH 主题词是由美国国立医学图书馆编制的

权威词表, 可用于 PubMed 中文章的检索. PubMed 为文章提供人工标注的 MeSH 主题词, 概括文章主题且准确性较高. 2014 版的 MeSH 共包含了 27 149 个 MeSH 主题词. 本文结合“特征项—文章”间的关系和“文章—实体”关系, 建立“特征项—实体”间的关系, 从而实现特征项对实体(基因和临床术语)的矢量表达.

1.2.2 方法过程

基于以上思想, 本方法选取所有 MeSH 主题词组成特征项集合, 矢量表达所有待研究的基因和临床术语. 本文借用经典 TF-IDF 思想计算特征项的权值, 此过程需考虑两个影响因素——一是 MeSH 主题词与该实体的相关程度, 二是 MeSH 主题词在文献集中的表现力, 方法如图 1 所示.

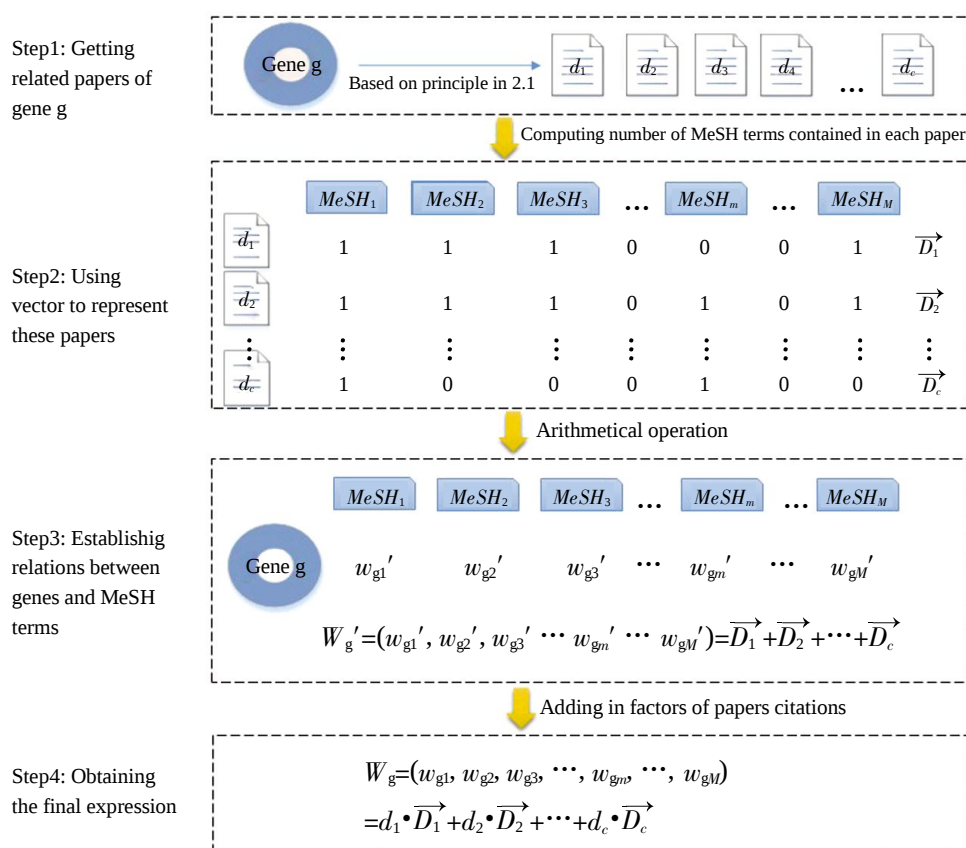


Fig. 1 Computational process of the MeSH-based method

This figure shows the calculation process of vector expression of gene g. d_i represent papers related to gene g, while c is the count of related papers. For each paper, there are different MeSH terms in PubMed annotated for them. M is total count of MeSH terms. For MeSH₁, 1 stand for that paper 1 contains this MeSH terms, while 0 represent the opposite meaning. In this way, paper 1 can be represented as $\vec{D}_1$. So we can establish associations between genes and MeSH terms according to combination of gene-papers association and papers-MeSH terms associations. Paper citations were taken into consideration finally, d_i is amended to the average number of citations.

对于“实体—文章”的关系, 本文主要通过 2.1 中描述的方法, 如基因 g, 获取其相关文章, 假设共 c 篇. 对于“文章—MeSH”的关系, 我们用特征项集合 MeSH(MeSH₁, MeSH₂, ..., MeSH_M) 矢量表达每篇文章, 1 表示该文章包含对应 MeSH 主题词, 0 表示未包含. 如图 1 中, 对于第 1 篇文章, 矢量表示结果 $\vec{D}_1 = (1, 1, 1, 0, \dots, 0, 0, 0, \dots, 0, 0, 1)$,

表示该篇文章包含 4 个 MeSH 主题词, 分别为 MeSH₁, MeSH₂, MeSH₃, MeSH_M. 结合“实体—文章”和“文章—MeSH”建立“实体—MeSH”之间的关系. 对于基因 g, 本文定义矢量 W'_g 表示为 $W'_g = \vec{D}_1 + \vec{D}_2 + \dots + \vec{D}_c$ 即所有相关文章矢量之和. 然而, 不同的文章被引用次数不尽相同, 著名算法 PageRank^[13] 引入网页相互指向数量对网页排名, 极

大地提高了搜索结果的准确性. 因此, 本文通过利用文章相互引用信息对矢量表达结果进行修正. d_i 表示第 i 篇文章平均每年被引用次数, 则修正后的矢量表示结果为 $W_g=(w_{g1}, w_{g2}, w_{g3}, \dots, w_{gm}, \dots, w_{gM})=d_1 \cdot \vec{D}_1+d_2 \cdot \vec{D}_2+\dots+d_c \cdot \vec{D}_c$. 综上所述, 算法数学计算过程如下: 对于基因 g , 对应于第 m 个 MeSH 特征项的权重 w_{gm} 表示为:

$$w_{gm}=T_{gm} \times IDF_m=T_{gm} \times \lg\left(\frac{N}{n_m}\right) \quad (1)$$

其中 T_{gm} 表示第 m 个 MeSH 特征项对应于基因 g 的相关程度, N 表示本研究的文献库中包含文章的总数, n_m 表示其中含有主题词 m 的文章数量. T_{gm} 计算方式如下:

$$T_{gm}=\frac{\sum_{j=1}^k d_i}{\sum_{j=1}^c d_j \times n_{gj}} \quad (k \leq c) \quad (2)$$

其中 c 表示基因 g 相关文章数量, k 表示其中包含 MeSH 主题词 m 的文章数量. 其中 d_i 表示第 i 篇与基因 g 相关的文章平均每年被引用次数, 共 k 篇, n_{gj} 表示与 g 相关的文章中, 第 j 篇文章对应的 MeSH 主题词的个数, d_j 与 d_i 类似.

由上一步得到每个 MeSH 特征项对应于基因 g 和临床术语 t 的权值后, 基因 g 可表示为 $W_g=(w_{g1}, w_{g2}, \dots, w_{gM})$, 临床术语 t 可表示为 $W_t=(w_{t1}, w_{t2}, \dots, w_{tM})$. 本文的方法采用数学中计算矢量余弦值的方法, 计算两者的相关性, 表示如下:

$$association[g][t]=\cos[W_g, W_t]=\frac{\sum_{i=1}^M w_{gi} \cdot w_{ti}}{\sqrt{\sum_{i=1}^M w_{gi}^2 \times \sum_{i=1}^M w_{ti}^2}} \quad (3)$$

由此, 对于每对临床术语 - 基因, 我们可以得到相应的余弦相似性值.

2 结果分析

为了验证以上方法, 本文选取结直肠癌作为研究对象. 结直肠癌是一种典型的癌症, 在我国恶性肿瘤死因顺位中居第五位^[14]. 相比于其他癌症, 结直肠癌具有相对清晰的分子机制, 利于检验研究结果. 此外, 在临床上, 结直肠癌的发现阶段对其治疗结果有很大影响. 本文选取与结直肠癌相关的 66 个临床术语作为起点, 利用上述方法分析它们与基因之间的余弦相似性值, 共获得 284 724 个 (66×4 314) 结果值.

本文首先使用接收者操作特征曲线(receiver operating characteristic curve, ROC 曲线)从数学模型角度评估方法性能, 并选取合适的阈值筛选结果, 用于结直肠癌的临床组学关系分析. 同时, 将本文提出的方法与 GeneCard 中经典 Novoseek 评分算法^[15]结果进行对比, 从实际应用角度评估方法的可靠性. 接着对得到的结果从基因、临床术语、关系三个角度分别进行分析和解释. 结合生物信息学工具 g:Profiler 和 KEGG 对基因的功能和通路进行生物学角度的理解和解释; 通过中心度等数学运算, 结合生物医学知识, 对临床概念进行分析; 结合“共现”文章数量, 对关系进行多角度的评估和阐述. 其整体流程如图 2 所示.

2.1 方法评估

本文主要从两个角度对基于 MeSH 主题词的方法进行评估. 利用操作特征曲线 ROC, 从数学角度评估方法的可靠性, 利用最佳工作点选取合适的阈值, 筛选结果; 与 GeneCards 中经典的 Novoseek 评分算法对比分析, 通过定量结果趋势评估本方法的可靠性.

a. ROC 曲线是一种坐标式的分析工具, 可用于对一个二分类器模型评估, 选择最佳分类模型; 并在同一个模型中设定最佳阈值. 利用 IBM SPSS 软件做 ROC 曲线分析. 得到 ROC 曲线的面积 AUC=0.86, 定量关系值为 0.6 时, 得到最大阳性似然比. 由此选取关系筛选阈值为 0.6, 最终保留余弦相似性大于 0.6 的关系, 用于进一步的分析中. 对应 36 个与结直肠癌相关的临床术语、203 个基因以及它们之间的 462 个关系.

b. 为了检验本文方法的可靠性, 与 GeneCards 数据库中经典的 Novoseek 评分算法进行对比分析. Novoseek 评分算法是一种基于“共现”原则的经典方法. 选取三个与结直肠癌高度相关的代表性基因, 即抑癌基因 APC、错配修复基因 MLH1、MSH2, 以及 GeneCards 中与这些基因相关的临床术语作为测试对象. 通过本文提出的基于 MeSH 主题词方法对测试对象进行定量分析, 得到结果与 Novoseek 评分算法结果对比如图 3 所示. 其中数据系列 P 代表本方法得到的结果, 数据系列 N 代表 GeneCards 中提供文献挖掘结果. 由于 Novoseek 评分算法中, 绝对数值并无实际意义, 因此这里只比较两组数据的趋势. 图中两组数据趋势一致性为方法可靠性提供了保证.

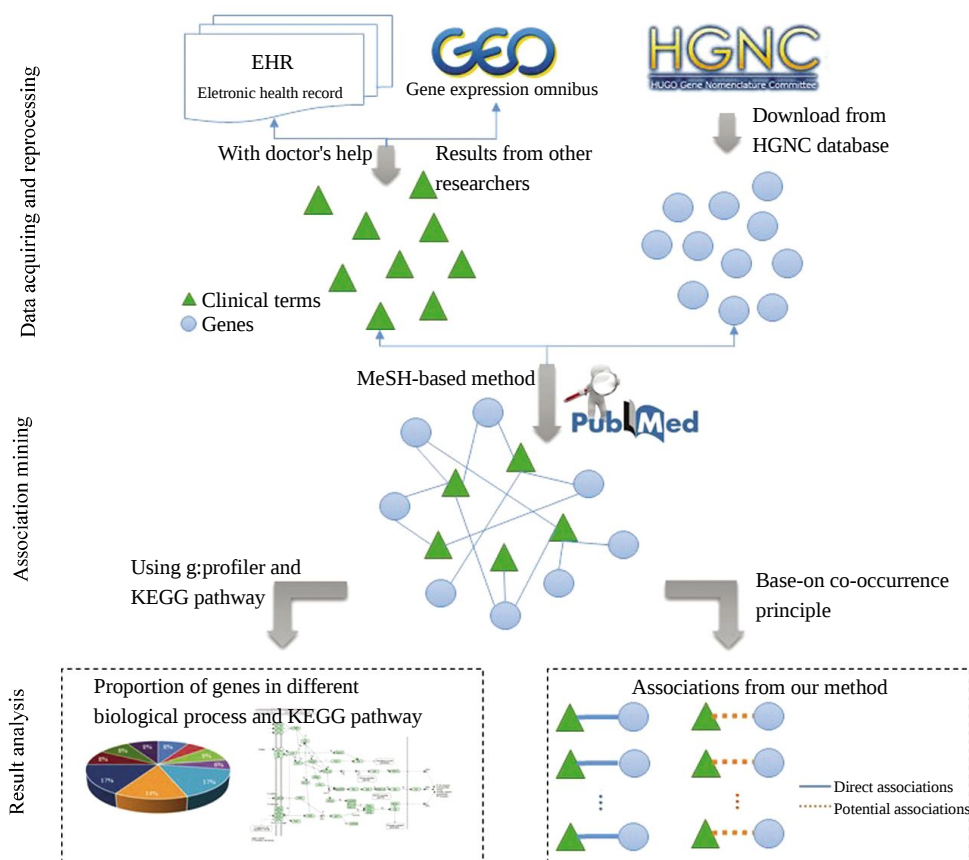


Fig. 2 Workflow of the MeSH-based method

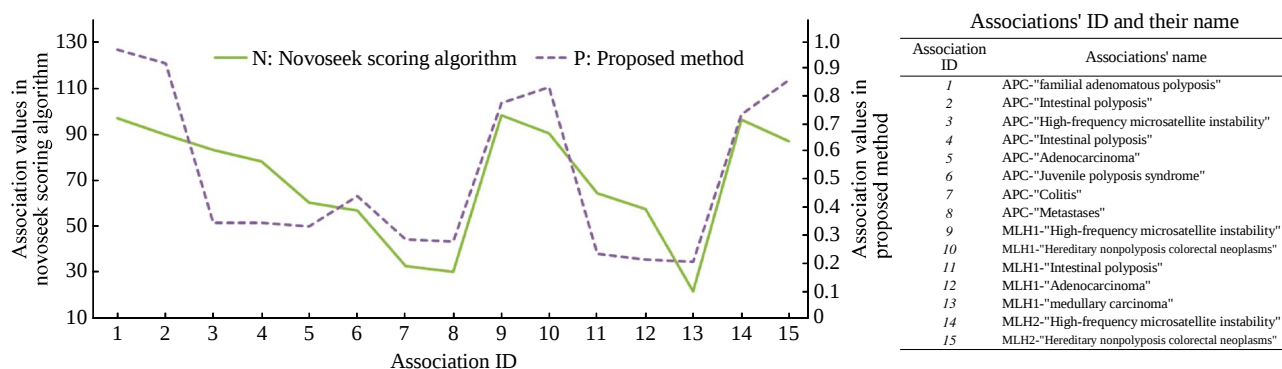


Fig. 3 The comparison of results from our MeSH-based method and GeneCards

The different data series on the horizontal axis represent the different associations between representative genes related to CRC and clinical terms. The series N represents the results from Novoseek scoring algorithm (NSK, abbreviation), while the series P represent the results of our MeSH-based method. In order to better compare their tendency, double vertical axis are used in this figure and the left one represents the NSK, while the right one refers to our MeSH-based method.

2.2 结合生物医学知识的分析

对于得到的 203 个与结直肠癌相关的基因, 本文结合生物信息学工具 g:Profiler 对其功能进行分

析. g:Profiler 用于大规模实验中获取基因功能信息, 可捕捉到基因在基因本体(Gene Ontology, 简称 GO)中的标注信息, 采用 Benjamini-Hochberg

统计学模型控制多次重复实验中假发现率的问题, 从而提高分析结果的准确性^[16]. 本文主要使用 g:Profiler 对基因所在基因本体中的生物过程

(biological process)进行分析, 统计每个通路和生物过程中基因所占的比例, 如图 4 所示.

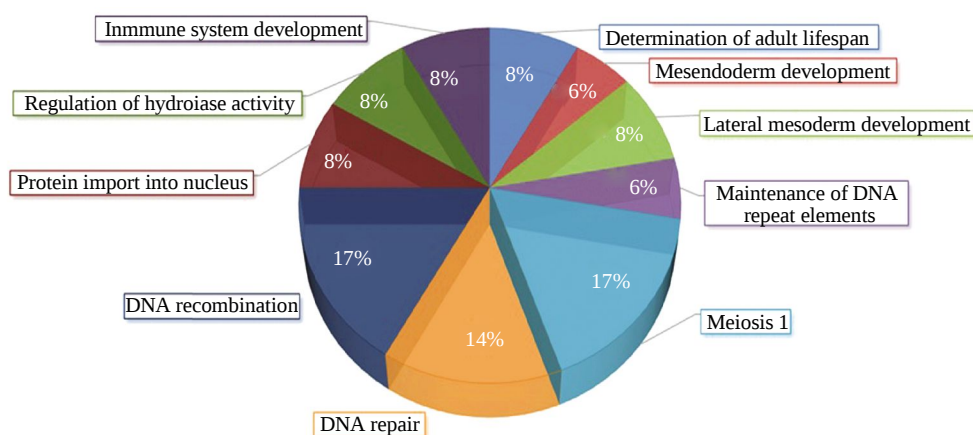


Fig. 4 Proportion of genes in different biological process

图 4 显示了挖掘出的基因在基因本体中已标注的, 在各个生物过程中的比例(未标注的将被自动过滤). 其中生物过程: 减数分裂 I (meiosis I)、DNA 修复 (DNA repair)、DNA 重组 (DNA recombination)包含基因比例分别为 17%、14%和 17%. 这些生物过程中基因的异常或缺陷容易导致疾病的发生, 比如 DNA 错配修复的缺陷容易导致微卫星不稳定, 增加其他基因的不稳定性, 进而出现整个基因组的不稳定, 导致细胞增殖和分化的异常, 促进肿瘤的发生.

基因与基因组的京都百科全书 (Kyoto Encyclopedia of Genes and Genomes, KEGG)^[17]是根据基因信息理解细胞或组织的上层功能性意义与用途的生物信息学资源. 本文使用 KEGG 中的 KEGG GENES 数据库和 KEGG PATHWAY 数据库

对获取的结直肠癌相关基因进行分析, 获得它们在各个 Pathway 中的分布情况, 表 1 是覆盖基因最多的 10 个通路及其中的代表性基因. 其中包括了结直肠癌通路(colorectal cancer)、错配修复通路(mismatch repair)和黏着斑通路(focal adhesion). 已有研究^[18]指出, 遗传非息肉性结直肠癌的特点是 DNA 错配修复(DNA mismatch repair)缺陷, 从而导致微卫星不稳定性, 这在临床中已作为肿瘤遗传不稳定的一个敏感检测指标. 另有文献表明^[19], 黏着斑激酶在结直肠癌的初级阶段和转移阶段都会过表达. 而在结直肠癌的通路中, 则包含了重要的 DNA 修复基因(MLH1、MSH2、MSH3、MSH6)和抑癌基因(APC)以及与细胞凋亡有关的 BAD 基因等.

Table 1 Top 10 KEGG pathways covering the largest number of input genes

Pathways number	Pathways name	Example genes
hsa05200	Pathways in cancer - Homo sapiens (human)	APC, LAMB1, MET, MLH1 <i>etc.</i>
hsa05210	Colorectal cancer - Homo sapiens (human)	APC, APC2, BAD, MLH1 <i>etc.</i>
hsa01100	Metabolic pathways - Homo sapiens (human)	DLST, DPYD, HAL <i>etc.</i>
hsa04550	Signaling pathways regulating pluripotency of stem cells - Homo sapiens (human)	BMPR1A, BMPR1B, NODAL <i>etc.</i>
hsa03430	Mismatch repair - Homo sapiens (human)	MLH1, MLH3, MSH2, MSH3, MSH6 <i>etc.</i>
hsa04390	Hippo signaling pathway - Homo sapiens (human)	TEAD1, APC2 <i>etc.</i>
hsa04514	Cell adhesion molecules (CAMs) - Homo sapiens (human)	ICAM2, CADM3 <i>etc.</i>
hsa04510	Focal adhesion - Homo sapiens (human)	LAMB1, BAD, JUN <i>etc.</i>
hsa05213	Endometrial cancer - Homo sapiens (human)	BAD, MLH1 <i>etc.</i>
hsa04060	Cytokine-cytokine receptor interaction - Homo sapiens (human)	MET <i>etc.</i>

本研究认为, 一个临床术语相关基因数量越多, 说明该临床术语在结直肠癌的研究中受关注程度越高, 两个临床术语共同相关的基因越多, 它们之间存在关联的可能性更大. 临床术语的中心度能反映出与结直肠癌的相关度. 表 2 列出中心度最高的前 10 个临床术语.

Table 2 Top10 clinical terms ranked by centrality

Rank	Clinical terms	Relation counts
1	Colorectal undifferentiated carcinoma	117
2	Metastases	68
3	Grade I colorectal adenocarcinoma	47
4	High-frequency microsatellite instability	25
5	Adenocarcinoma	21
6	Abdominal mass	19
7	Body weight decreased	15
8	Tenesmus	12
9	Perineural invasion	11
10	Diabetes mellitus	11

由表 2 可知, “未分化癌”和“转移瘤”是中心度最高的两个临床术语, 中心度分别为 117 和 68. 未分化癌是一种高度恶性的上皮性肿瘤, 不具有任何腺管结构, 癌细胞可呈单一性或多形性, 属于高级别癌^[20]. 而结直肠癌转移瘤是指结直肠癌肿瘤细胞从原发部位侵入淋巴管、血液或其他途径被带到它处继续生长, 形成与原发部位肿瘤相同类型的肿瘤. 结直肠癌转移可涉及肝、肺、肾上腺等重要器官. 未分化癌和转移瘤均属结直肠癌后期表现, 这与相关基因数目多是一致的. 此外, 我们发现, 糖尿病(diabetes mellitus)也在其中. 已有研究表明^[21], 糖尿病患者患有结直肠癌的危险度是非糖尿病患者的 1.72 倍. 对于糖尿病是否一定会引起结直肠癌, 或者糖尿病人出现哪些症状或哪些基因发生突变时, 一定会导致结直肠癌, 还有待进一步的研究.

本文统计得到的前 50 个关系对应的共现文章数量(按照 2.1 所述方法, 统计临床术语 - 基因对出现文章数量), 得到其中 80%的关系对应共现文章数量大于 0. 表 3 列出其中前 10 个关系及其“共现”文章数量.

Table 3 TOP10 gene-clinic associations in our results

RelationID	Genes	Clinical terms	CosValue	Count of co-occurrence papers
1	APC	Familial adenomatous polyposis	0.9728	1467
2	MTA1	Perineural invasion	0.9387	0
3	COLCA2	Colorectal undifferentiated carcinoma	0.9059	0
4	COLCA1	Colorectal undifferentiated carcinoma	0.9030	0
5	IBD2	Colitis	0.8917	21
6	NODAL	Perineural invasion	0.8858	23
7	IBD4	Colitis	0.8857	3
8	IBD3	Colitis	0.8823	16
9	NODAL	Lymphatic metastases	0.8757	620
10	IBD3	Ulcerative colitis	0.8744	15

在得到的前 10 个关系中, 发现其中 3 个“共现”文章数为 0, 分别是关系 2, 关系 3 和关系 4. 在关系 2 中, 嗜神经性侵袭(PNI)是指肿瘤细胞侵袭和转移到神经系统, 发生 PNI 是肿瘤恶化的一个标志, 也是恶性肿瘤低治愈率, 高死亡率的标志. PNI 可以在远离于肿瘤扩散源的部位, 或者没有淋巴及血管的部位发生, 侵袭程度超出局部肿瘤侵袭的程度. 已有研究表明^[22], 在多种癌症中, 基因 CDH1(E-cadherin)表达的缺失与上皮间质转化、

癌症的浸润及转移均有关. 另一项研究指出, 在前列腺癌中, MTA1 通过对 ATK 的激化, 调控 CDH1 的表达. 为本文推测 MTA1 与嗜神经性侵袭的关系, 提供可靠的参考. 基因 COLCA1 和 COLCA2 的全称分别是 colorectal cancer associated 1 和 colorectal cancer associated 2. COLCA1 基因编码一种跨膜蛋白定位于颗粒状结构, 包括晶体嗜酸性颗粒等颗粒状细胞器. 已经被证明是结直肠癌的易感基因. 未分化癌是一种高度恶性的上皮性肿

瘤, 属于高级别瘤。目前没有文章同时在摘要或标题中提到两者与未分化癌的关系, 进一步的实验研究有助于关系的验证和分析。本文还分析了前 50 个关系中, “共现” 文章数量为 0 的临床 - 组学关系。其中包括 IBD9-Crohn Disease、MSH4-Hereditary nonpolyposis colorectal neoplasms(遗传非息肉性结直肠癌, HNPCC)、MTA1-Perineural invasion 等。Crohn Disease(克罗恩病)^[23]是一种慢性透壁性炎症病变, 可累及消化道中的任何部位, 并且可以引起胃肠外表现。2%~22% 克罗恩病患者的一级亲属患有炎症肠炎。对于家族性的炎症性肠病患者全基因组测序, 定位 9 个易感基因位点(IBD1~9), 根据本论文结果推测, IBD9 可能成为克罗恩病的易感基因。HNPCC 被认为是由于错配修复基因的种系突变导致的, 目前已被识别的认为与 HNPCC 有关的基因有: MLH1, MSH2, MSH6, PMS1, PMS2, MSH3。而 MSH4 也是 DNA 错配修复 mutS 家族中的成员之一, 虽然它并不参与 DNA 错配修复, 但已有研究表明^[24]它与 MSH5 可能参与有丝分裂 DNA 修复和损伤应答。该研究还指出, MSH4 很有可能独立参与除减数分裂以外的生物过程。这些研究为 MSH4 与 HNPCC 存在潜在关系提供证据, 为基础实验验证提供了有力的数据。

3 总 结

本文基于 PubMed 中的 MeSH 主题词提出一种关系定量挖掘的方法, 借用信息领域中“向量空间模型”思想, 结合 TF-IDF 算法和余弦相似性, 用 MeSH 主题词矢量表示待挖掘对象, 并考虑了文献间的相互引用关系对矢量表达的影响, 最后通过矢量间余弦相似性的计算, 定量分析疾病与基因之间的关系。方法借用“向量空间模型”的思想, 有效地将多样生物实体关系的分析转化为简单的数学运算, 扩大待分析文章的范围; 人工标注的 MeSH 主题词作为特征项集合, 避免复杂的语法和语义分析, 同时, 弥补了向量空间模型对语义敏感度欠佳的不足; 此外, 方法避免“共现”原则要求“同时出现”的限制, 可实现直接关系和潜在关系的挖掘。本文利用结直肠癌疾病作为研究对象, 利用该方法对与结直肠癌相关的临床术语和基因之间的相互关系进行挖掘, 结合生物医学知识的分析和解释, 提出 IBD9-Crohn Disease、MSH4-Hereditary nonpolyposis colorectal neoplasms、MTA1-Perineural invasion 等潜在关系, 供临床研究和鉴定。本文从

生物信息学角度, 结合计算机和数学方法, 实现了海量生物医学文献库中临床 - 组学关系的挖掘和分析。分析结果对于基础医学和分子机理的研究(如生物通路的研究、标志物的发现等)提供生物信息学分析依据, 有助于促进基础研究者对疾病机理的认识和探索。为更好地促进研究方法和研究结果的应用, 我们计划后期提供“数据共享平台”, 实现数据的查询、下载以及方法的共享, 同时有利于基础研究者对方法的再次利用, 提高数据分析的针对性。此外, 在方法上, 我们将通过人工筛选和基于规则的方式, 对 MeSH 特征项集合进行精炼, 进一步提高数据的准确性。本论文选取结直肠癌作为研究对象, 该方法也可用于其他疾病的分析和研究中, 为大量文献数据中潜在知识的挖掘提供一种新的方法和思路。

附件 表 S1, 本文中使用的与结直肠癌相关的所有临床术语, 见本文网络版附录(<http://www.pibb.ac.cn>)。

参 考 文 献

- [1] Siegel R L, Miller K D, Jemal A. Cancer statistics, 2015. CA: A Cancer Journal for Clinicians, 2015, **65**(1): 5-29
- [2] Weinberg R. The Biology of Cancer. New York: Garland Science, 2013: 32-69
- [3] Deng N, Zheng L, Liu F, *et al.* CrcTRP: a translational research platform for colorectal cancer. Computational and Mathematical Methods in Medicine, 2013, 2013
- [4] The PubMed database [EB/OL]. <http://www.ncbi.nlm.nih.gov/pubmed/>
- [5] Özgür 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-i285
- [6] Narayanasamy V, Mukhopadhyay S, Palakal M, *et al.* TransMiner: mining transitive associations among biological objects from text. Journal of Biomedical Science, 2004, **11**(6): 864-873
- [7] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, **18**(11): 613-620
- [8] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, **27**(8): 861-874
- [9] Sayers E. Entrez programming utilities help. URL <http://www.ncbi.nlm.nih.gov/books/NBK25499>, 2009
- [10] Liu F, Feng Y, Li Z, *et al.* Clinic-genomic association mining for colorectal cancer using publicly available datasets. BioMed Research International, 2014, **2014**: 170289
- [11] Povey S, Lovering R, Bruford E, *et al.* The HUGO gene nomenclature committee (HGNC). Human Genetics, 2001, **109**(6): 678-680

- [12] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, **24** (5): 513–523
- [13] Page L, Brin S, Motwani R, *et al.* The PageRank citation ranking: bringing order to the Web. 1999: 1–6
- [14] Kushi L H, Doyle C, McCullough M, *et al.* American Cancer Society guidelines on nutrition and physical activity for cancer prevention. *CA: A Cancer Journal for Clinicians*, 2012, **62** (1): 30–67
- [15] Safran M, Dalah I, Alexander J, *et al.* GeneCards Version 3: the human gene integrator. *Database*, 2010, **2010**: baq020
- [16] Reimand J, Kull M, Peterson H, *et al.* g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 2007, **35** (suppl 2): W193–W200
- [17] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 2000, **28**(1): 27–30
- [18] Jass J R. Pathology of hereditary nonpolyposis colorectal cancer. *Annals of the New York Academy of Sciences*, 2000, **910** (1): 62–74
- [19] Lark A L, Livasy C A, Calvo B, *et al.* Overexpression of focal adhesion kinase in primary colorectal carcinomas and colorectal liver metastases immunohistochemistry and real-time PCR analyses. *Clinical Cancer Research*, 2003, **9**(1): 215–222
- [20] 郑树, 万德森. 结直肠癌. 北京: 北京大学医学出版社, 2008: 32–60
- Zheng S, Wan D S. *Colorectal Cancer*. Beijing: Peking University Medical Press. 2008: 32–60
- [21] Hu S Q, Tang Z, Zhang M. Investigation on the risks of colorectal cancer in patients with diabetes mellitus. *World Chinese Journal of Digestology*, 2007, **15**(1): 88
- [22] Fan L, Wang H, Xia X, *et al.* Loss of E-cadherin promotes prostate cancer metastasis *via* upregulation of metastasis associated gene 1 expression. *Oncology Letters*, 2012, **4**(6): 1225–1233
- [23] Phillips, Robin KS, Sue Clark, *eds.* *Colorectal Surgery: Companion to Specialist Surgical Practice*. Elsevier Health Sciences, 2013: 1–7
- [24] Wu X, Her C, Xu K. The Role of MutS Homologues MSH4 and MSH5 in DNA Metabolism and Damage Response. *INTECH Open Access Publisher*, 2011: 1–9

A MeSH-based Biomedical Literature Mining Method for Exploring Associations Between Genes and Clinical Terms*

FENG Ya-Ning, JIAO Meng-Ying, DUAN Hui-Long, DENG Ning**

(College of Biomedical Engineering and Instrument Science, Key Laboratory of Biomedical Engineering of Ministry of Education of China, Zhejiang University, Hangzhou 310027, China)

Abstract The causes and progressions of cancers have close associations with the mutations of genes in our body, which lead to abnormal symptoms and detection indicators. Therefore, providing clinical decision support for early diagnosis and precise treatment of cancers is very urgent and necessary, which can be achieved by mining the associations between genes and clinical behaviors from conclusive biomedical literature data. A MeSH-based (Medical Subject Headings, MeSH) method was proposed for biomedical objects association mining in this paper. By using MeSH (which is provided in PubMed) to represent each object as a vector in the Vector Space Model and taking the citations between articles into consideration, we translated the associations mining into mathematical operating successfully. We finally obtained 203 genes and 462 associations related to colorectal cancer (CRC) after applying our method in the associations mining between genes and clinical behaviors of CRC. In order to analyze and verify the mining results, some bioinformatics tools, such as g:Profiler and KEGG were used for functions and pathway analysis of genes. The results show that this MeSH-based method works robust in the association mining. Besides removing the restriction of co-occurrence for the indirect associations mining, our proposed method also avoid complex grammatical analysis which lead to massive calculation.

Key words literature mining, quantitative analysis, association mining, colorectal cancer, PubMed, medical subject headings

DOI: 10.16476/j.pibb.2015.0129

*This work was supported by grants from The National Natural Science Foundation of China(31100592), Hi-Tech Research and Development Program of China (2012AA02A601) and National Science and Technology Major Project of China(2013ZX03005012).

**Corresponding author.

Tel: 86-571-87951792, E-mail: zju.dengning@gmail.com

Received: April 29, 2015 Accepted: July 15, 2015

附录

Table S1 Colorectal cancer related clinical terms used in this paper

TermNo.	Clinical term
1	Abdominal mass
2	Abdominal pain
3	Adenocarcinoma
4	Adenomatous polyposis coli
5	Adenosquamous carcinoma
6	Albumins
7	Anemia
8	Appendicitis
9	Ascites
10	Asthenia
11	Backache
12	Blood glucose
13	Blood urea nitrogen
14	Body weight decreased
15	CA-125 antigen
16	Carcinoembryonic antigen
17	Carcinoid tumor
18	Carcinoma <i>in situ</i>
19	Cholecystitis
20	Colitis
21	Colon carcinoma metastatic to the liver
22	Colorectal adenosquamous carcinoma
23	Colorectal medullary carcinoma
24	Colorectal signet ring cell carcinoma
25	Colorectal small cell carcinoma
26	Colorectal squamous cell carcinoma
27	Colorectal undifferentiated carcinoma
28	Constipation
29	Creatinine
30	Crohn Disease
31	Diabetes mellitus
32	Diarrhea
33	Dyspepsia
34	Eosinophilia
35	Familial adenomatous polyposis
36	Fecal occult blood
37	Grade I colorectal adenocarcinoma
38	Hematochezia
39	Hemoglobin
40	Hereditary nonpolyposis colorectal neoplasms
41	High-Frequency microsatellite instability
42	Intestinal obstruction
43	Intestinal polyposis

Continued

TermNo.	Clinical term
44	Intestinal polyps
45	Juvenile polyposis syndrome
46	Local neoplasm recurrence
47	Lymphatic metastases
48	Lymphocyte count abnormal
49	Medullary carcinoma
50	Metastases
51	Mucinous adenocarcinoma
52	Nausea
53	Neutrophils decreased
54	Papillary adenocarcinoma
55	Papillary carcinoma
56	perineural invasion
57	Platelet count measurement
58	Schistosomiasis
59	Signet ring cell carcinoma
60	Small cell carcinoma
61	Squamous cell neoplasms
62	Tenesmus
63	Thymidylate synthase
64	Ulcerative colitis
65	Vomiting
66	White blood cell count abnormal