肿瘤表观遗传学研究热点的聚类分析

李晓丽(湖北医药学院附属人民医院,湖北 十堰 442000)

摘　要:[目的] 以 Web of Science 中的 SCIE 数据库为数据来源,分析肿瘤表观遗传学热点研究问题以及研究趋势。[方法] 从 SCIE 数据库收录的肿瘤表观遗传学研究文献提取了相关的题录信息和研究关键词,利用 SPSS18.0 软件进行共词聚类分析,提取高频主题词,生成共现矩阵、共词聚类。[结果] 2003~2007 年的聚类前 5 类类团分别为肿瘤表观遗传、继承、DAN 甲基化、表观遗传过程和组蛋白修饰;而 2008~2012 年的聚类前 7 类类团分别肿瘤表观遗传、非编码 RNA、DAN 甲基化、组蛋白修饰、低甲基化、表观遗传过程和组蛋白乙方酰化,说明表观遗传、低甲基化与肿瘤表观遗传、肿瘤表观遗传过程以及组蛋白修饰与肿瘤表观遗传已经成为肿瘤表观遗传学研究领域的研究热点, 也是肿瘤表观遗传学研究学者比较关注的内容。[结论] 肿瘤表观遗传学的研究目前除了对肿瘤表观遗传过程、印记丢失等方面进行分析之外,主要侧重于肿瘤表观遗传学的形成原因分析,具体包括低甲基化与肿瘤的发生、组蛋白修饰与肿瘤的发生以及非编码 RNA 与肿瘤的发生研究。肿瘤表观遗传学的研究正由原来单一化、表面化的研究逐渐转向深层次、复杂化的研究。

关键词:肿瘤表观遗传学;高频主题词;共现矩阵;聚类分析
中图分类号:R730.23　　**文献标识码:**A　　**文章编号:**1004-0242(2013)04-0284-04

Clustering Analysis on Hot Research on Tumor Epigenetics

LI Xiao-li
(Shiyan People's Hospital Affiliated to Hubei University of Medicine, Shiyan 442000, China)

Abstract: [Purpose] To analyze the hot research issues and its research trend of tumor epigenetic. [Methods] Key words and related bibliographic information on tumor epigenetic research literatures were selected from SCIE database and analyzed by SPSS18.0 software for co-word clustering; then high frequency MeSH wre extracted to generate the co-occurrence matrix and co-word clustering. [Results] The top 5 clustered word groups from 2003 to 2007 were tumor epigenetics, inheritanc, DNA methylation, epigenetic processes and histone modification respectively; and the top 7 clustered word groups from 2008 to 2012 were tumor epigenetics, non-coding RNAs, DNA methylation, histone modification, hypomethylation, epigenetics processes and histone acetylation respectively. That showed epigenetics, hypomethylation and tumor epigenetics, tumor epigenetics processes and histone modification and tumor epigenetics had become the focus in the field of tumor epigenetics research and the concern of the tumor epigenetics researchers. [Conclusion] Researches on tumor epigenetics at present are mainly focused on analysis of reasons for the formation of tumor epigenetics, specifically on hypomethylation and tumorigenesis, histone modification and tumorigenesis, and non-coding RNAs and tumorigenesis in addition to researches on tumor epigenetic processes, loss of impriting and so on. Tumor epigenetics research is gradually turning from the simplified, superficial research to the deep and complicated one.

Key words: tumor epigenetics; high-frequency; co-occurrence matrix; clustering analysis

表观遗传学是一门探索从基因到表型变化的学科,研究不仅涉及 DNA 序列改变基因的相关表达,同时也涉及基因调控遗传因素等方面的表达[1]。表观遗传学将传统医学意义上的遗传学与环境因素进行结合,增加了研究的复杂性,同时也使人们从一个全新的角度去研究生命现象。表观遗传的发生时间并没有限制,可能存在于胚胎时期,同时也可能存在于成体阶段。表观遗传信息在细胞减数分裂的过程中传递给下一代, 不再随着蛋白质基因编码序列的改变而改变[2]。体内的遗传信息在特定的基因表达条件下启动或者停止,进而进行基因表达。正是由于基因表达的选择性导致细胞表型的各不相同。目前表观遗传通常被定义为基因表达通过有丝分裂或减数分裂发生了可遗传的改变, 而 DNA 序列不发生改变。

收稿日期:2012-10-26;修回日期:2012-12-10
E-mail:sysrmyytsg@163.com

表观遗传学在生命科学中是一个普遍而又十分重要的研究领域[3]。随着表观遗传学在各领域中的应用，其在肿瘤疾病的防治中有着十分广泛的应用，具有良好的发展前景。对肿瘤表观遗传学的研究现状以及进展进行研究，将有助于肿瘤表观遗传学研究领域的进一步深化。本文将 SCIE 数据库作为论文研究的基础数据来源，利用数据挖掘方法，主要利用 SPSS 软件对相关数据进行整理分析，对肿瘤表观遗传学的研究热点以及研究趋势作出分析判断。

1 资料与方法

1.1 数据来源

选取 SCIE 数据库为数据来源，检索与肿瘤表观遗传学相关的研究文献。研究过程中将研究文献的时间锁定于 2003 年 1 月 1 日至 2012 年 8 月 20 日。由于 10 年的数据量比较庞大，而且 10 年期间的研究动向前后相差较大，因此论文将基本数据分为 2003~2007 年和 2008~2012 年两组。将 Tumor and Epigenetic 作为检索的主题词，共检索出文献 65 618 篇，其中 2003~2007 年共 22 257 篇，2008~2012 年共 43 361 篇。

1.2 研究方法

采用 Excel 对相关数据进行处理，然后采用 SPSS 软件对 2003~2007 年和 2008~2012 年两个时间段的肿瘤表观遗传学关联高频词篇进行系统聚类分析。在聚类分析的过程中，主要用 Ochiai 系数作为聚类情况的表征因素。论文的研究分析过程主要包括以下两个方面：

①词篇矩阵与共现矩阵的生成：在论文的前期研究过程中，利用 Excel 对相关数据进行整理分析，利用书目共现系统对文献题录中的主题词进行提取。文献主题词的提取主要利用帕欧公式进行提取，旨在提取主要主题词，避免不相关词的影响。

②共词矩阵的生成：论文主题词共词矩阵的生成主要采用 SPSS 分析软件的聚类分析方法。聚类分析方法的分析原理是文献主题词在标引的过程中会出现多个主题词反映文献内容情况，而检索出的主题词又会在多篇文章中出现。如果出现了上述共词矩阵情况，就可以认为这些主题词之间关系较为紧密，进而可以判断该领域的研究情况。

2 结 果

2.1 高频主题词判定

通过对第一步研究过程中的文献题录数据信息进行分析，对 2003~2007 年、2008~2012 年的相关文

Table 1 The high frequency MeSH (from 2003 to 2007)

No.	High frequency MeSH	N	%	Cumulative percentage(%)
1	Epigenetic,tumor	780	8.0233	7.0231
2	Inheritanc	652	6.2321	14.2554
3	DNA methylation	321	3.4656	17.7210
4	Epigenetic processes	297	3.1412	20.8622
5	Histone modification	176	2.9014	23.7636
6	Hypomethy1ation	253	2.9001	26.6637
7	Histone acetylation	253	2.5014	29.1651
8	Island methylator phenotype	245	2.3612	31.5263
9	Methyltransferas	212	2.1718	33.6981
10	Cgp Island	201	2.1315	35.8296

Table 2 The high frequency MeSH (from 2008 to 2012)

No.	High frequency MeSH	N	%	Cumulative percentage(%)
1	Epigenetic,tumor	993	8.2231	8.2231
2	Inheritanc	613	4.2405	12.4636
3	DNA methylation	548	4.4645	16.9281
4	Epigenetic processes	425	3.0435	19.9716
5	Histone modification	413	2.9812	22.9528
6	Hypomethy1ation	408	2.9024	25.8552
7	Histone acetylation	318	2.5232	28.3784
8	Island methylator phenotype	287	2.3416	30.7200
9	Methyltransferas	278	2.1734	32.8934
10	Cgp Island	214	2.1236	35.0170

Table 3 The high frequency MeSH co-occurrence matrix (2003~2007)

MeSH	Epigenetic	Inheritanc	DNA thylation	Epigenetic processes
Epigenetic,tumor	780	553	231	215
Inheritanc	553	652	145	187
DNA methylation	231	145	321	165
Epigenetic processes	215	187	165	297

Table 4 The high frequency MeSH co-occurrence matrix (2008~2012)

MeSH	Epigenetic	Inheritanc	DNA thylation	Histone modification
Epigenetic,tumor	993	513	465	378
Inheritanc	513	613	315	246
DNA methylation	456	315	548	378
Histone modification	378	246	378	425

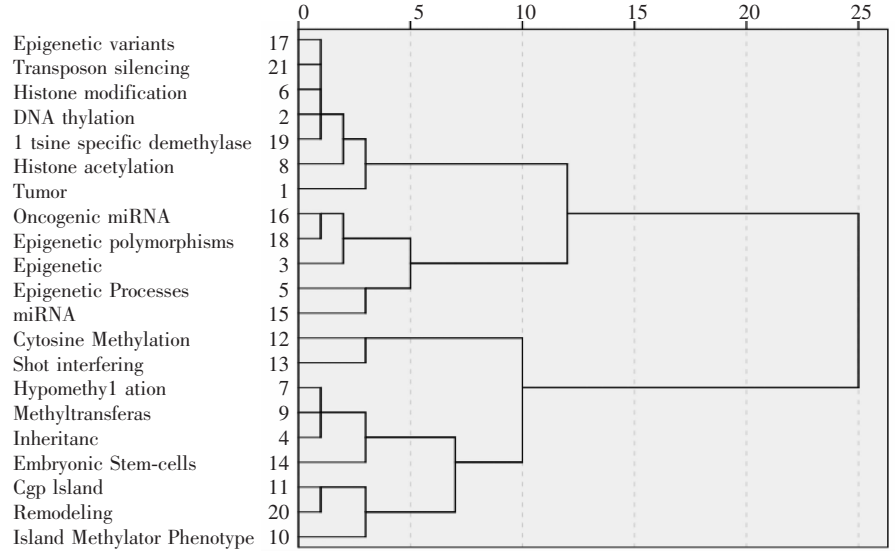

Figure 1 The literature high frequency MeSH dendrogram in 2003~2007

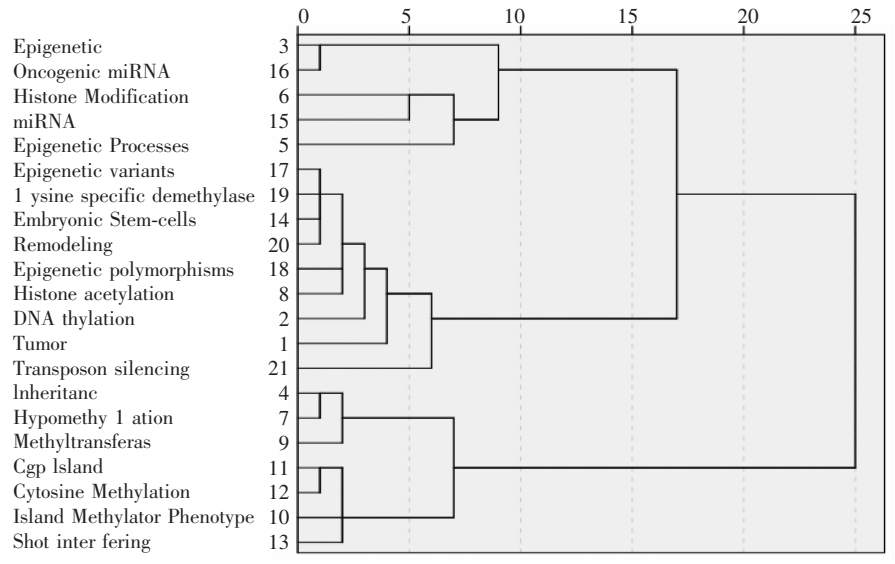

Figure 2 Document high frequency MeSH dendrogram in 2008~2012

献主题词进行分析可以发现:2003~2007 年文献主题词个数为 3 390 个，主题词中出现次数为 1 的有 2 430 个;2008~2012 年的相关文献主题词为 4 497 个,主题词中出现次数为 1 的有 1 983 个。根据帕欧公式和实际数据分布情况，确定主题词中最低频值为 55 以上的 24 个词为第一组高频主题词，应用相同的方法,第二组高频主题词为 23 个。由于高频主题词的数量偏多,论文的研究过程中仅选取前 10 位高频主题词为研究分析对象(Table 1、2)。

2.2 共现矩阵

对 2003~2012 年期间两组数据的高频主题词进行共现矩阵分析。由于高频主题词数量较多,而论文的研究目的是分析肿瘤表观遗传学的研究进展,因此论文选择位列前 4 位的高频主题词进行共现矩阵分析(Table 3、4)。

2.3 系统聚类结果

利用 SPSS18.0 对 2003~2012 年文献的词篇矩阵进行聚类分析,将文献高频主题词生成聚类树图(Figure 1、2)。

将 2003~2007 年和 2008~2012 年两组数据的高频文献主题词进行梳理,将两个时间段内的聚类情况、高频主题词情况进行对比分析，可见 2003~2007 年，聚类类团前 5 位分别为肿瘤表观遗传(Epigenetic,tumor)、继承(Inheritanc)、DAN 甲基化(DNA methylation)、表观遗传过程(Epigenetic processes)和组蛋白修饰(Histone modification)。而 2008~2012 年聚类类团前 7 位分别为肿瘤表观遗传(Epigenetic,tumor)、非编码 RNA、DNA 甲基化(DNA methylation)、组蛋白修饰(Histone modification)、低甲基化(Hypomethylation)、表观遗传过程(Epigenetic processes)和组蛋白乙酰化(Histone acetylation)。2003~2007 年第 1、3、4、5 类团和 2008~2012 年的第 1、2、3、6 类团相同。

3 讨 论

3.1 聚类对比分析

2003~2007 年的第 1、3、4、5 类团分别与 2008~

2012 年的第1、2、3、6 具有相同的含义，说明表观遗传、低甲基化与肿瘤表观遗传、肿瘤表观遗传过程以及组蛋白修饰与肿瘤表观遗传已经成为肿瘤表观遗传学研究领域的研究热点，也是肿瘤表观遗传研究学者比较关注的对象。在 2003~2007 年，除了上述 4 种研究主题词之外，还包括了继承这个研究主题词。在 2008~2012 年，除了上述 4 种主题词外，还包括低甲基化、非编码 RNA 和组蛋白乙酰化 3 个关键词。这也表明学术界对肿瘤表观遗传学的研究在逐渐深入化和全面化。究其主要的原因在于最近几年以来，表观遗传学在肿瘤的防治方面取得了很好的进展，为了解决实际过程中遇到的复杂问题，需要对其形成原因进行深入分析，这也是肿瘤表观遗传学研究越来越深入的原因。从 2003~2007 年期间的研究主要关键词可以看出这个阶段的肿瘤表观遗传学研究侧重于形成原因分析，但分析的程度不够深入。从 2008~2012 年这个阶段的研究文献可以看出肿瘤表观遗传学应用研究和理论研究进入了更深层次的发展阶段，其研究热点主要侧重于低甲基化、组蛋白修饰、非编码对肿瘤表观遗传的影响。总体来说，论文对数据库中研究数据的统计分析结果与肿瘤表观遗传的研究现状基本吻合。

3.2 聚类消失分析

2003~2007 年主题词中的继承关键词在 2008~2012 年的主要研究关键词中不再出现。继承方面的研究主要出现在肿瘤表观遗传学中的“印记丢失”研究。肿瘤表观遗传学异常的提出就源于对印记丢失的研究，印记丢失主要是遗传印记的丢失。所谓遗传印记，是指来自父母双方的等位基因在通过精子和卵子传递给子代时发生了修饰，使带有亲代印记的等位基因具有不同的表达特性[4]。肿瘤中一些印记的丢失就会导致异常情况的发生。随着肿瘤表观遗传学研究的不断深化，表观遗传与肿瘤之间的关系研究越来越细化，而不是简单地用继承关系来概括，这也是在本文的数据分析中出现继承关键词聚类消失的情况，数据分析结果也符合肿瘤表观遗传学研究动向。

3.3 聚类新增分析

通过对 2008~2012 年的数据分析可以发现，研究文献中新增了非编码 RNA、低甲基化和组蛋白乙酰化 3 个类团。这也说明非编码 RNA、低甲基化和组蛋白乙酰化很可能或者已经成为肿瘤表观遗传学的研究热点。非编码 RNA 是指不编码蛋白质的 RNA，其中包括 rRNA，tRNA，snRNA，snoRNA 和 microRNA 等多种已知功能的 RNA，还包括未知功能的 RNA[5]。这些 RNA 的共同特点是都能从基因组上转录而来，但是不翻译成蛋白质。例如，近年来医学研究中发现遗传性非息肉性结直肠癌的发生就与 miRNA 有着直接关系。组蛋白乙酰化会引发肿瘤发生。染色质重塑可导致核小体位置和结构的变化，引起染色质变化，进而导致肿瘤的发生。从上述聚类分析对比分析结果可以看出，肿瘤表观遗传学的研究热点主要集中于低甲基化、组蛋白修饰、非编码 RNA 与肿瘤发生之间的关系，同时还涉及肿瘤表观遗传过程和组蛋白乙酰化等研究。

综上，对 2003~2007 年和 2008~2012 年的两组文献聚类类团进行分析可以得出以下结论：①肿瘤表观遗传学的研究目前除了对肿瘤表观遗传过程、印记丢失等方面进行分析之外，主要侧重于肿瘤表观遗传学的形成原因分析，具体包括低甲基化与肿瘤的发生、组蛋白修饰与肿瘤的发生以及非编码 RNA 与肿瘤的发生研究；②肿瘤表观遗传学的研究正逐渐由原来单一化、表面化的研究转向深层次、复杂化的研究。肿瘤表观遗传学的应用研究在日渐强化。例如甲基化检测可以作为肿瘤非常有效的指标，对 miRNA 的表达定量检测，使 miRNA 成为肿瘤诊断的有效的标志物，这些实际应用案例说明了肿瘤表观遗传学的应用研究越来越深化。肿瘤表观遗传学的研究遵循理论研究不断深化、实践应用研究不断强化的发展思路。

参考文献：

[1] Kato Y，Sasaki H.Imprinting and looping: epigenetic marks control interactions between regulatory elements[J]. Bioessays，2005，27(1):1-4.

[2] Frigola J，Song J，Stirzaker C，et al.Epigenetic re-modeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band[J].Nature Genetics，2006，38(5):540-549.

[3] Sartore-Bianchi A，Fieuws S，Veronese S. Standardisation of EGFR FISH in colorectal cancer: results of an international interlaboratory reproducibility ring study[J]. J Clin Pat-hol，2012，65(3):218-223.

[4] Lv F，Shao ZY，Xie ZL. Epigenetics in cancer diagnosis and treatment [J].Western medicine，2008，20(1): 196-198.[吕锋，邵泽勇，谢朝良.表观遗传学在肿瘤诊治中的应用[J]. 西部医学，2008，20(1)：196-198.]

[5] Ren LL，Fang JY.Autophagy in tumor epigenetics research progress [J].Chinese Journal of Cancer，2011，21 (6): 484-488.[任琳琳，房静远.肿瘤中自噬的表观遗传学研究进展[J].中国癌症杂志，2011，21(6)：484-488.]