中药指纹图谱相似度计算方法探析

关洪月[1,2],李林[1,2],刘晓[1,2],殷放宙[1,2],蔡宝昌[1,2,,3*]
(1. 南京中医药大学药学院,南京 210046; 2. 南京中医药大学国家教育部中药炮制规范化及标准化工程研究中心,南京 210029; 3. 南京海昌中药集团,南京 210061)

[摘要] 通过检索 CNKI,SCIENCE DIRECT 数据库,对近年来出现的中药指纹图谱相似度计算方法进行收集和整理,筛选出 20 篇相关文献,共共总结出 9 种相似度计算方法,包括峰重叠率法(Nei 系数法)和峰重叠率与共有峰强度结合法(改进 Nei 系数法)、距离系数法、相关系数和夹角余弦法以及在此基础上作出的改进的其他 6 种方法。通过分析,各种方法均有各自的优点和缺点,并分析其各自的特点和不足在实际应用过程中应根据需要,以便在实际应用过程中采用恰当的相似度评价方法,使指纹图谱技术更好地应用于中药质量控制。

[关键词] 中药;指纹图谱;相似度;计算方法
[中图分类号] R284 [文献标识码] A [文章编号] 1005-9903(2011)18-0282-06

Study on Similarity Algorithm of Traditional Chinese Medicine Fingerprints

GUAN Hong-yue[1,2], *LI Lin*[1,2], *LIU Xiao*[1,2], *YIN Fang-zhou*[1,2], *CAI Bao-chang*[1,2,3*]
(1. *College of Pharmacy, Nanjing University of Chinese Medicine, Nanjing* 210046, *China*; 2. *Engineering Center of State Ministry of Education for Standardization of Chinese Medicine Processing, Nanjing University of Chinese Medicine, Nanjing* 210029, *China*; 3. *Nanjing Haichang Chinese Medicine Group Co. Ltd.*, *Nanjing* 210061, *China*)

[**Abstract**] To summarize the similarity algorithm methods used in the analysis of fingerprints for traditional Chinese medicine in recent years through CNKI, SCIENCE DIRECT database retrieval. Twenty articles were found andNinenine kinds of similarity calculation methods were summed up including Nei coefficient and Improved Nei

[收稿日期] 20110309(007)
[基金项目] 国家科技部重大专项(2009ZX09308-004)
[第一作者] 关洪月,在读硕士,从事中药指纹图谱研究,Tel:025-86798281, E-mail:guanhongyue2099@163.com
[通讯作者] *蔡宝昌,Tel:025-85811112,E-mail:bccai@126.com

关系的探讨[J]. 中国中医药科技, 2003, 10(2): 65.
[2] 胡晓文. 大剂量复方丹参注射液抗肝纤维化作用临床观察[J]. 甘肃中医学院学报, 1996, 13(1):24.
[3] 叶红军. 丹参和白细胞介素 2 防治大鼠免疫性肝纤维化的实验研究[J]. 中华消化杂志, 1994, 14(5):266.
[4] 马学惠. 丹参对实验性肝硬化细胞外基质影响的免疫组化观察[J]. 肝脏病杂志, 1994, 2(2):79.
[5] 杨卫东. 丹参的氧自由基清除作用[J]. 中国药理学通报, 1990,2: 118.
[6] 和水祥. 丹参对培养人胚肝细胞胶原合成的影响[J]. 中国医学科学院学报, 1996, 18 (1):70.
[7] 马红. 黄芪对免疫损伤性肝纤维化大鼠的治疗作用[J]. 中西医结合肝病杂志, 1997, 7(1):32.
[8] 何燕,胡志峰,李平,等. 柴胡皂苷抗肝纤维化大鼠脂质过氧化作用的研究[J]. 中国中药杂志, 2008,33(8):915.
[9] 陈爽, 贲长恩, 杨美娟,等. 柴胡皂甙对肝细胞增殖及基质合成的实验研究[J]. 中国中医基础医学杂志, 1999,5(5):21.

[责任编辑 邹晓翠]

coefficient, distance coefficient, correlation coefficient, cosine of angle and other six improving similarity algorithm methods. The advantages and the advantagesand disadvantages of each method were listed. In practical work, similarity methods should be chosen according to their characteristics, in which way this technique would be applied to the quality control of TCM more appropriately.

[**Key words**] traditional Chinese medicine; fingerprint; similarity; algorithm

近年来,指纹图谱技术由于具有适合从整体上分析复杂化学物质组成的稳定性的特点,已成为国内外广泛接受的中药质量评价模式,许多国家均已接受指纹图谱用来评价中药质量。美国 FDA 在有关草药质量控制中提出可以建立并申报药材及其制剂的色谱指纹图谱,以考察草药产品批间质量的一致性。欧共体在草药质量指南中也指出,草药的质量稳定性仅靠测定已知有效成分是不够的,应通过指纹图谱显示其所含的多种成分。英国草药典、印度草药典、德国药用植物协会、加拿大药用及芳香植物学会等均把指纹图谱作为中药质量控制的标准之一。

要形成实用的指纹图谱,首先要收集大量合格的样本,通过色谱法或者光谱法等建立指纹图谱,在此基础上选择合适的方法,形成共有模式,后续待检测的样本则是通过一定的计算方法计算出与共有模式的相似度,通过相似度来评价中药质量的真假优劣。对于合格样本的收集和共有模式的形成研究均比较深入,而相似度的计算方法的研究仍有待加强。

1 中药指纹图谱相似度计算方法

1.1 峰重叠率法(Nei 系数法)与峰重叠率与共有峰强度结合法(改进的 Nei 系数法)

Nei 系数法是较早提出的相似度计算方法,通过比较各图谱峰数的相似程度判断药材之间的相关性。公式为:

$$r = \frac{2n_0}{n_1 + n_2} \times 100\% \quad (1)$$

张聪等[1]运用 Nei 系数法对 11 批红参甲醇提取液的指纹图谱计算重叠率,通过对八强峰的比较分析,初步表明国产红参和进口高丽红参品质十分接近,可以等同使用。但此方法只是计算了各图谱峰数相似程度,没有考虑到峰强度,目前已基本不单独使用。

孟庆华等[2]在此基础上提出了改进 Nei 系数法,原理是通过将同一条件下建立的各中药指纹图谱中各个峰的相对保留值按从小到大顺序进行编号,若某一指纹图谱在某相对保留值处无峰(相对峰面积或相对峰高为零),仍给相应的编号,以保证各色谱指纹图谱都有相同的色谱峰数。以相对峰面积或者相对峰高为纵坐标,以相对保留值或者编号为横坐标,建立二维体系,将同一张指纹图谱的所有点连接起来构成色谱指纹特征曲线。通过比较特征曲线的相似程度来判断中药的真假、优劣。此方法结合了峰强度信息,且无须对相对保留值数据进行标准化等预处理,但改进后仍对小峰缺失敏感而相对大峰不够敏感[3]。公式为:

$$f = \frac{2n_0}{n_1 + n_2} - \frac{2}{n_1 + n_2} = \sum \left| \frac{X_{ik} - X_{jk}}{X_{ik} + X_{jk}} \right| \quad (2)$$

n_0 是两图谱共有峰数,以下用 n 表示;n_1,n_2 分别为待测图谱与标准图谱组成峰总数。X_{ik},X_{jk}分别代表第 i 个样品和第 j 个样品的第 k 个特征峰值,即峰高或峰面积。

1.2 距离系数法 距离系数法常用于描述样品间的亲疏程度,系数越大,两者差异越大。距离系数种类较多,本文以欧氏距离 d_{ij} 和马氏距离 D^2 为例。

$$d_{ij} = \sqrt{\left[\sum_{k=1}^{n} (X_{ik} - X_{jk})^2 \right]} \quad (3)$$

$$D^2 = (X_{ik} - X_{jk})^T s^{-1} (X_{ik} - X_{jk}) \quad (4)$$

(s^{-1}为样本协方差矩阵 s 的逆矩阵)

欧氏距离系数法以几何中两点间的距离大小,对应指纹图谱即以两图谱相应峰面积大小的绝对差异(不区分大小峰差异)来反映两图谱间的相似性,在一定范围内无论是大峰还是小峰都表现出较高的敏感性[3],在某种程度上适用于与总量有关的中药与中药材的质量分析。曹建军等[4]用 HPLC 指纹图谱技术检测生地黄炮制成熟地黄过程中主要活性成分梓醇、5-羟甲基糠醛(5-HMF)、麦角甾苷等及指纹的变化,并通过欧氏距离系数法与标准熟地黄指纹图谱比较,最终确定生地黄的最佳炮制时间为 26 h。

马氏距离的原理与欧氏距离类似,通常与化学模式识别如聚类分析、主成分分析联合运用以评价药材质量。吴昊等[5]采用聚类分析及马氏距离判别的多元统计学方法对参麦注射液的 HPLC 指纹图谱中采集的数据进行分析,通过聚类分析将 18 个样本分为 3 类:工艺 A 为 1 类,工艺 B 为 1 类,伪品为 1 类。选取聚为 1 类的工艺 A 的 9 个样品作为合格样品,通过计算其他待测样品与该合格品库的马氏距离来判断样品是否合格。

这 2 种距离系数法各有优缺点。中药各成分常常作为协同作用的整体达到治疗疾病的目的,欧氏系数并没有考虑各指纹峰的相关性和方差的差异性,同时数值有单位量,计算会受单位以及响应值差异等影响[6];马氏距离与欧氏距离不同的是它考虑到了指纹峰各种特性之间的联系,不受量纲的影响,但是它夸大了变化微小的变量的作用。距离系数法虽然具有一定程度的定量判别能力,但无法直接给出直观的综合定量评价结果。

1.3 相关系数与夹角余弦法 这 2 种方法均是将每张指纹图谱看作是一个由图谱组成峰相应积分面积值 $x_1, x_2, x_3, \cdots, x_n$ 构成的一个 n 维空间向量 $\bar{X} = (x_1, x_2, x_3, \cdots, x_n)$,相关

系数法在数学中是用来比较 2 个数据集合是否在同 1 条直线上,在指纹图谱中通过比较各向量的相关系数 r_{ij}来比较样品的相似程度,夹角余弦法则是比较各向量之间的夹角余弦值 $\cos\theta$ 来反映样品的相似程度。王龙星等[7]利用向量夹角法,对 11 个不同产地及炮制方法的吴茱萸样品的液相色谱指纹图谱进行相似度计算,结果此法能较好地评价指纹图谱间的相似性,并清楚地区别了汤洗 7 遍这种炮制方法对指纹图谱的影响。田润涛等[8]采用计算机拟合技术结合灯盏细辛药材色谱指纹图谱对不同相似度评价方法的性能和指纹图谱特征指标的选择标准进行了考察。结果表明在以色谱峰面积作为指标进行的相似度计算中,推荐采用夹角余弦和相关系数法。

这 2 种方法是目前应用最多的 2 种相似度算法,但这 2 种方法是两图谱间形状的测度,与图谱相应峰比例无关,即如果同一样品按照比例放大或缩小 n 倍时,2 种方法计算得到的相似度仍为 1,这 2 种算法不能反映在样品含量上的相似程度。这 2 种方法对峰缺失的敏感性与改进 Nei 系数法相反,对大峰变化敏感,相对小峰不敏感[3]。对于同属不同种的药材,大峰的差异是有可能的,此时采用相关系数法、夹角余弦法评价相似度还是较为适宜的,但不适合不同批次的中药制剂质量稳定性的控制。

$$r_{ij} = \frac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)^2}\sqrt{\sum_{k=1}^{n}(X_{jk} - \bar{X}_j)^2}} \quad (5)$$

$$\cos\theta = \frac{\sum_{k=1}^{n}X_{ik}X_{jk}}{\sqrt{\sum_{k=1}^{n}X_{ik}^2}\sqrt{\sum_{k=1}^{n}X_{jk}^2}} \quad (6)$$

1.4 程度相似度法、改良程度相似度法与新改良程度相似度法 20 世纪 90 年代,周美立[9-10]提出了用于计算系统相似度的理论。这一相似系统理论用系统组成的要素及其特征来表述系统的相似度,即将系统相似度表述为相似要素的多少及其相似程度大小的函数,并给出数学模型和系统相似度的多元函数数学表达式 $Q = f(K, L, n, u)$。其中,K 和 L 分别表示 2 个系统的各自组成要素的数量,n 表示相似要素的数量,u 表示相似要素的相似程度。在系统 A 和 B 的 n 个相似要素中,对应的相似要素组成相似元。在系统要素一定的条件下,对于系统相似度优劣的评价,既要看相似要素的多少,也要看每一相似要素的相似程度。由相似要素的数量确定的相似度为数量相似度 Q_n;由相似元值的大小确定的相似度为程度相似度 Q_u。基于相似系统理论,刘永锁等[11-12]提出了衡量两色谱指纹图谱相似系统间的程度相似度 Q 和改良程度相似度 Q'计算方法。1 张色谱指纹图谱可看做为 1 个相似系统,每个色谱峰是该相似系统的相似要素,两图谱对应的色谱峰组成相似元,每个相似元对应的特征值为对应峰峰面积的比值。

$$Q = \frac{1}{n}\sum_{k=1}^{n}\frac{\min(X_{ik}, X_{jk})}{\max(X_{ik}, X_{jk})} \quad (7)$$

$$Q' = 1 - \frac{1}{n}\sum_{i=1}^{n}|1 - X_{ik}/X_{jk}| \quad (8)$$

通过计算机模拟数据以及实验数据表明以相关系数、夹角余弦作为相似度指标,对样品的差异不灵敏。有些数据经标准化预处理之后计算结果也没有多大的改善,经对数转换预处理之后还会改变数值的特性,因此对于中药的质量控制,不适宜采用这种数据预处理的方法。以 10 批栀子药材提取物为例,分别采用相关系数、夹角余弦、程度相似度和改良程度相似度与参照样品比较,结果显示相关系数和夹角余弦计算得到的相似度都在 0.99 以上,反应不出 10 批样品与参照样品的差异,而基于相似系统理论的程度相似度和改良程度相似度直接针对样品差异计算,均能反映出 10 批样品与参照品的差异。这 2 种方法在一定程度上解决了相关系数和夹角余弦对数据差异不灵敏问题。

詹雪艳等[13]基于标准偏差比平均偏差更能突现大偏差的思想在改良相似度基础上进行了改进,提出了新改良程度相似度 q'。在共有峰峰面积差异不超过 100% 的情况下,q'计算方法能够更灵敏地反映两样本的峰面积的相对差异,能突出组分群中某些组分对既定配比的较大偏离,适用于中药成分群配比波动的控制和中药过程质量控制。

$$q' = 1 - \sqrt{\frac{\sum_{k=1}^{n}\left(1 - \frac{X_{ik}}{X_{jk}}\right)}{n}} \quad (9)$$

这 3 种计算方法存在一定的缺陷,样品与参照的差异不大时,以程度相似度计算,结果反映出的差异与相对差异接近,随着样品与参照的相对差异增大,计算结果反映出的差异与实际差异的误差也变大。对于峰面积的差异不超过 100% 时,改良相似度与新改良程度相似度均可以准确反映样品与参照的差异,q'更灵敏一点,然而当部分峰的相对差异远 >100% 时,并且这些峰对相似度计算结果起主导作用时,Q'和 q'会出现负值而无意义。

1.5 全定性全定量相似度 孙国祥等[14]在向量夹角余弦相似度的基础上提出了比率定性相似度。即是将对照指纹向量 $\bar{Y} = (y_1, y_2, \cdots, y_n)$作 $P_0 = (1, 1, 1, \cdots, 1)$处理,同理,样品指纹向量 $\bar{X} = (x_1, x_2, \cdots, x_n)$作 $P_s = (x_1/y_1, x_2/y_2, \cdots, x_n/y_n) = (r_1, r_2, \cdots, r_n)$,则 P_s 与 P_0 间夹角余弦值就是比率定性相似度 S_F'。S_F 和 S_F'相差越小越好,<5% 时最理想。夹角余弦相似度 S_F 和比率定性相似度 S_F'构成全定性相似度。同时提出了模长百分比与宏观含量相似度,投影含量相似度 C% 与定量相似度 P %,含量相似度与平均质量百分数,校正含量相似度与校正平均质量百分数,分别构成第一、第二、第三、第四级全定量相似度。全定性全定量相似度分别从化学成分分布相似性和含量相似性 2 个方面评价样品与对照指纹图谱的相似程度。对样品质量进行评价时,全定性相似度均 >0.9 为必要条件,上述 4 种全定量相似度可

选择任意1组,推荐使用第二级全定量相似度,制剂控制在90% ~110%,原料控制在85% ~120%,组内相差不得超过10%为合格。全定性相似度和全定量相似度一方面可保证削减大指纹峰影响,等权对待小指纹峰贡献;另一方面,从突出大指纹峰对体系的作用出发进行评价,这样能同时兼顾检测所有指纹峰对体系的定性定量的贡献作用。全定性相似度和全定量相似度的密切结合构成指纹图谱新的质控体系,是利用指纹图谱宏观控制中药质量的另一比较好的方法。目前,已将这一评价方法已在多种中药的指纹图谱评价中得到运用[15-16]。

$$S_F = \cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{10}$$

$$S_F' = \frac{\sum_{i=1}^{n} r_i}{\sqrt{n\sum_{i=1}^{n} r_i^2}} \tag{11}$$

$$C\% = \frac{XL}{|\vec{Y}|} \times 100\% = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2} \times 100\% = SF \times W\% \tag{12}$$

$$P\% = S_F \times R\% \tag{13}$$

1.6 权值非均一性相似度 杨忠民等[17]为了减小夹角余弦法或相关系数法在指纹图谱分析中的误差,提出了峰面积权值非均一性进行相似度的计算,在此笔者称之为权值非均一性相似度S。标准样品与待测品所有峰平均值为($\sum X_i + \sum Y_i$)/2n,标准样品与待测品一个峰平均值为($X_i + Y_i$)/2,两者相除得到当前峰的权重即$\frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}{n(X_i + Y_i)}$,此时是将各峰分配了相同的权值。但在实际运算过程中,面积小的峰稍有轻微的差异,在最后的计算中表现的过于突出,这与现实情况不相符,此时由于把其权重设置的过大,应把面积较小的峰的权重适当调小;而面积大的峰有时有较大的差异性,在最后的计算中差异性表现的偏小,这与现实情况也不相符,此时权值定为相同不是很合适,应适当地把权重调大。为解决权值完全一致性,将上文的权重开方即$\sqrt{\frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}{n(X_i + Y_i)}}$。将总体差异性相对于标准品峰面积与待测样品峰面积之和($\sum |X_i - Y_i|$)/($\sum X_i + \sum Y_i$)乘以权重的开方则得到待测样品与标准样品的总体差异值。此算法中对权重开方,若相应二峰平均面积 < 总平均面积的则权重 >1,也就是面积较小的有较大的权重,开方可适当调小,其大小仍然 >1,仍兼顾面积小有效性可能高的这一事实;若相应二峰平均面积 > 总平均面积的则权重 <1,也就是面积较大的有较小的权重,故开方可适当调大,其大小仍然 <1,仍兼顾面积大有效性可能相对低的这一事实。夹角余弦法、相关系数法与数据线性比例没有关系,化学成分比例不变,但其总量变化,则二者难以体现这种变化;新算法能够体现这种成分比例的变化,这一方法适合于与总量有关的中药与香精香料的质量分析。

$$S = 1 - \frac{\sum_{i=1}^{n} \frac{|X_i - Y_i|}{\sqrt{X_i + Y_i}}}{\sqrt{n\left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i\right)}} \tag{14}$$

1.7 乘方相似度和定量乘方相似度 孙国祥[18]针对样品各峰X_i与对照指纹图谱相应峰的变化Y_i之间的积分值比较,提出乘方相似度S_i,公式如下。S_i较好地反映了X_i/Y_i百分变化关系,据此对所有指纹峰的S_i取平均值得到从整体上评价样品与对照指纹图谱间的定性相似度指标S_{gx}。S_i能够定性描述指纹成分与对照指纹图谱的差别,同时具有很好的定量描述功能,当$S_{gx} \geqslant 0.73$,成分含量在80 % ~ 120 %,$S_{gx} \geqslant 0.80$,成分含量在85 % ~115 %,$S_{gx} > 0.85$,成分含量在90 % ~110 %,$S_{gx} > 0.92$,成分含量在95 % ~105 %。但是乘方相似度在评价$x_i/y_i > 100\%$时,相似度变化 <1.0,实质是以1.0为中心发生变化,如x_i/y_i = 80 %和120 %的乘方相似度都是0.725。针对这一情况提出了定量乘方相似度S_{gxs}的概念。S_{gxs}能较好地解决此问题,它能够在总体上反应样品含量变化。将乘方相似度和定量乘方相似度应用于银杏叶提取物HPLC指纹图谱的评价,结果它们能够从宏观角度定性定量评价样品与对照指纹图谱间的相似程度,将得到的结果用宏观含量相似度R %和平均质量百分数M %进行对比验证,定量乘方相似度非常好地与R %,M %吻合。

$$Si = 5^{-Zi} = 5^{-\frac{Xi}{Yi}-1} \tag{15}$$

$$S_{gx} = \frac{1}{n}\sum_{i=1}^{n} 5^{-Zi} = \frac{1}{n}\sum_{i=1}^{n} 5^{-\frac{Xi}{Yi}-1} \tag{16}$$

$$S_{gxs} = \frac{1}{n}\sum_{i=1}^{n} 5^{Zi} \times 100\%$$

$$[Z_i = \frac{x_i}{y_i} - 1;\text{若 } x_i > y_i, Z_i = \frac{1}{2}\left(\frac{x_i}{y_i} - 1\right)] \tag{17}$$

1.8 夹角余弦法 杨云等[19]在对中药指纹图谱全谱计算夹角余弦系数的基础上,对指纹图谱的特征峰计算共有峰的折线重合率,经过加权重计算而得到更为准确的指纹图谱相似度值。此算法分两步进行,首先选择指纹图谱样本的所有点,设定有2个样本分别为X_i和X_j,m和n分别是样本X_i和X_j的最大采样数,设$p = \min(m,n)$,全谱夹角余弦比较公式表示为S_{ij},见公式18。其次,求得两图谱特征峰曲线的重合系数f_{ij},即上文提到的改进Nei系数。在得到样本1和样本2全谱比较的夹角余弦系数S_{ij},计算出样本的特征折线的重合系数f_{ij},可以通过相关的权重校正得到样本1和样本2的最终相似度S_{ij}^0。α和β的选取可以根据待测的中药

指纹图谱的不同而调整。对 3 组田基黄中药样本的指纹图谱原始信号和校正后的信号采用此方法计算相似度,实验结果表明了该方法的可行性和优越性。全谱夹角余弦法相对于单纯的特征峰夹角余弦法可以很好的反应图谱全局相似程度,但是缺乏对于特征峰比较的细节描述,从而丢失了相似度比较中最为重要的信息。本算法在全谱夹角余弦比较的基础上,采用特征峰折线的重合率作为校正,在选择共有峰时带阈值自动校正时间漂移,从而可以得到与实际情况更为符合的相似度比较结果,而且此算法在计算机编程中很容易实现。但是,该算法的普遍适用性和对处理更多种类中药图谱的稳定性,还有待进一步的深入研究和实验验证。

$$S_{ij} = \frac{\sum_{i=1}^{p} X_i X_j}{\sqrt{\sum_{i=1}^{p} (X_i)^2 \sum_{i=1}^{p} (X_j)^2}} \quad (18)$$

$$S_{ij}^0 = S_{ij} \times \alpha + f_{ij} \times \beta \quad (19)$$

1.9 假设检验 Feng Gan 等[20]定义了一个差异向量 $\overrightarrow{r}$ ($\overrightarrow{r} = \overrightarrow{X} - \overrightarrow{Y}$),接着提出了一个假设,$H_0(\bar{r} = 0)$;$H_1(\bar{r} \neq 0)$($H_0$ 表示两图无明显差异,H_1 表示有明显差异)($\bar{r} = \frac{1}{n}\sum_{i=1}^{n} ri$)。采用 t 检验,$t = |\bar{r}| \sqrt{n}$,若 $t > t_{\alpha,f}$(α 为置信度,默认为 0.05,f 为自由度),两图谱存在明显差异。结合贝叶斯理论用于假设检验。由 H_0,H_1 的后验概率公式可得到:

$$p(H_0 | \bar{r}) = \frac{p(\bar{r} | H_0)p(H_0)}{p(\bar{r} | H_0)p(H_0) + p(\bar{r} | H_1)p(H_1)} \quad (20)$$

$$p(H_1 | \bar{r}) = \frac{p(\bar{r} | H1)p(H_1)}{p(\bar{r} | H_0)p(H_0) + p(\bar{r} | H_1)p(H_1)} \quad (21)$$

其中 $p(H_0)$,$p(H_1)$ 为先验概率,一般情况下可以合理假设 $p(H_0) = p(H_1) = 0.5$,则有 $\frac{p(H_0 | \bar{r})}{p(H_1 | \bar{r})} = \frac{p(\bar{r} | H_0)}{p(\bar{r} | H_1)}$。为了简化分析过程,假设 H_1 中 $\bar{r} = 1$,由于 $p(\bar{r} | H_0)$ 趋向于 $N(0, 1/n)$,$p(\bar{r} | H_1)$ 趋向于 $N(1, 1/n)$,得到 $\rho = \frac{p(H_0 | \bar{r})}{p(H_1 | \bar{r})} = e^{(n/2)(2\bar{r}-1)}$。若 $\rho > 1$,两图谱无显著差异。

以 3 个产地的川芎为例,分别是四川,江西和广东。每个产地样品共进样 10 次,分 2 天进样,每天各产地样品进样 5 次。用夹角余弦和相关系数法计算同一产地的样品相似度均在 0.999 左右,而这与 PCA 图上显示的四川和江西 2 个产地的样品均存在明显分界的结果不一致。从 PCA 图上可以看出四川和江西 2 个产地 10 批药材根据两天被明显分为 2 类(第 1 天为 1 类,第 2 天为 1 类),但用夹角余弦和相关系数计算得到的值相似度均很高,导致很难明确区分同一产地药材的相似程度或者差别。以上述假设检验方法分别对 3 个产地的川芎进行分析,结果表明,同一产地的川芎即使是同一天的样品不同批次之间还是存在差别,第 2 天的样品并不是所有的批次都与第 1 天有差别。

此方法以差异向量 $\overrightarrow{r}$ 的均值作为假设检验的统计量,通过 t 检验结合贝叶斯假设检验可以很好反应药材不同批次之间的差异,但是由于贝叶斯假设检验是建立在先验概率均为 0.5 的情况下,此方法仍存在不足之处,先验概率如何合理取值将是未来研究的一个重要问题。

2 结语

目前相似度计算方法已被开发成软件在指纹图谱研究中得到广泛使用。中南大学开发的软件采用相关系数和夹角余弦作为相似度的评价指标,浙江大学开发的软件采用夹角余弦作为相似度的评价指标,西北大学开发的相似度软件采用相关系数、夹角余弦、模糊分布以及欧氏距离系数法 4 个评价指标。相似度评价软件的出现大大简化了计算过程。国家药典委员会推荐使用中南大学和浙江大学开发的指纹图谱相似度计算软件。纵观以上几种相似度算法,每一种指纹图谱相似度算法都有其自身的特点和适用范围,在中药质量控制的过程中,只有针对不同评价方法制定相应合理的评价指标,并结合化学模式识别如聚类分析、主成分分析以及人工神经网络等,整个评价研究才对中药质量控制具有实际意义。中药指纹图谱的研究在国内仍属起步阶段,今后应加强指纹图谱与药效相关性的研究,还应加速指纹图谱研究和应用的产业化推广,从而推动我国中药产业现代化研究进程。

[参考文献]

[1] 张聪, 王智华, 金德庄. 中国红参与高丽红参的指纹谱(HPLC-FPS)比较研究[J]. 中成药, 2001, 23(3): 160.

[2] 孟庆华, 刘永锁, 王健松, 等. 色谱指纹图谱相似度的新算法及其应用[J]. 中成药, 2003, 25(1): 4.

[3] 聂磊, 曹进, 罗国安, 等. 中药指纹图谱相似度评价方法的比较[J]. 中成药, 2005, 27(3): 249.

[4] 曹建军, 梁宗锁, 杨东风, 等. 应用 HPLC 指纹图谱技术确定熟地黄炮制终点[J]. 中国中药杂志, 2010, 35(19): 2556.

[5] 吴昊, 田燕华, 郭平平, 等. 多元统计学在参麦注射液指纹图谱中的应用[J]. 中成药, 2002, 24(1): 3.

[6] 谷瑞敏, 涂洪谊, 孙鹤. 中药色谱指纹图谱相似度计算方法的探讨[J]. 中成药, 2009, 31(7): 988.

[7] 王龙星, 肖红斌, 梁鑫淼, 等. 一种评价中药色谱指纹谱相似性的新方法:向量夹角法[J]. 药学学报, 2002, 37(9): 713.

[8] 田润涛, 谢培山. 色谱指纹图谱相似度评价方法的规范化研究(一)[J]. 中药新药与临床药理, 2006, 17(1): 40, 54.

[9] 周美立, 王浣尘. 相似系统的分析与度量[J]. 系统工程, 1996, 14(4): 1.

[10] 周美立, 吴报任. 系统相似度数值方法[J]. 安徽工学院学报, 1991, 10(3): 67.

[11] 刘永锁, 孟庆华, 蒋淑敏, 等. 相似系统理论用于中药色谱指纹图谱的相似度评价[J]. 色谱, 2005, 23(2): 158.
[12] 刘永锁, 曹敏, 王义明, 等. 相似系统理论定量评价中药材色谱指纹图谱的相似度[J]. 分析化学研究报告, 2006, 34(3): 333.
[13] 詹雪艳, 史新元, 展晓日, 等. 基于相似系统理论的相似度计算方法的改进[J]. 分析化学研究简报, 2010, 38(2): 253.
[14] 孙国祥, 宋杨, 毕雨萌, 等. 色谱指纹图谱全定性相似度和全定量相似度质控体系研究[J]. 中南药学, 2007, 5(3):263.
[15] 孙国祥, 刘金丹, 侯志飞, 等. 全定性全定量相似度法评价甜瓜蒂的毛细管电泳指纹图谱[J]. 中南药学, 2007, 5(6): 558.
[16] 孙国祥, 侯志飞, 张春玲, 等. 色谱指纹图谱定性相似度和定量相似度的比较研究[J]. 药学学报, 2007, 42(1): 75.
[17] 杨忠民, 李忠民, 赵日利, 等. 指纹图谱相似度新算法的研究[J]. 中国测试技术, 2008, 34(3): 141.
[18] Guoxiang S, Jindan L. Qualitative and quantitative assessment of the HPLC fingerprints of *Ginkgo biloba* extract by the involution similarity method[J]. Anal Sci, 2007, 23: 955.
[19] 杨云, 朱学峰. 一种新的计算中药指纹图谱相似度方法与实现[J]. 计算机测量与控制, 2007, 15(10): 1376.
[20] Feng G, Runyi Y. New appoach on similarity analysis of chromatographic fingerprint of herbal medicine[J]. J Chromatogr A, 2006, 1104:100.

[责任编辑　邹晓翠]