

DOI: 10.3724/SP.J.1096.2010.00253

基于相似系统理论的相似度计算方法的改进

詹雪艳 史新元 展晓日 王 耘 乔延江*

(北京中医药大学中药学院, 北京 100102)

摘 要 针对常用的相似度方法对色谱指纹图谱数据的相对差异响应不灵敏, 本文对基于相似系统理论的改良程度相似度的计算方法进行了改进。模拟数据和乳块消片色谱指纹图谱数据计算结果表明, 改进后的新改良程度相似度 q_c 可区分与参照样本的总差异相同, 但差异分布不同的样本。在共有峰峰面积差异不超过 10% 的情况下, 新改良程度相似度 q_c 计算方法能够更灵敏地反映两样本的峰面积的相对差异。本研究的新改良程度相似度 q_c 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。

关键词 相似系统理论; 相似度算法; 过程质量控制

1 引 言

用色谱指纹图谱的相似度控制中药质量的稳定均一, 是中药质量控制的有效手段之一。色谱指纹图谱常用的相似度有夹角余弦、皮尔森相关系数和欧式距离等计算方法, 但是相关系数和夹角余弦相似度不能灵敏地反映小峰的变化和丢失, 不适合不同批次的中药制剂质量稳定性的控制。近年出现的新相似度计算方法: 比率定性相似度^[1]、总量统计矩相似度^[2-3]、加权马氏距离相似度^[4], 以及用相对熵的歧异值来衡量色谱指纹图谱的相似度^[5]等方法。文献[6-7]等用模拟数据与实验数据探讨了相关系数和夹角余弦对样本矢量数据差异的灵敏性问题, 提出了基于相似系统理论的程度相似度和改良程度相似度, 一定程度上可解决相关系数和夹角余弦对小色谱峰反应不灵敏的问题。然而, 如果两种样本相对于参照样本的总相对差异相同, 而差异分布不同, 如矢量中各个元素具有相近差异的样本和矢量中某些元素具有大的相对差异的样本, 改良程度相似度对其难以区分。针对以上问题, 本研究基于标准偏差比平均偏差更能突现大偏差的思想, 对改良相似度进行了改进, 改进后的新改良程度相似度突出样本中元素的大差异, 可较好地反映样本间的相对差异。

2 改良程度相似度的改进和数据来源

2.1 基于相似系统理论的改良程度相似度

20 世纪 90 年代, 周美立等^[8-9]提出了相似系统理论, 基于相似系统理论, 刘永锁等^[6-7]提出了衡量两色谱指纹图谱相似系统间的程度相似度和改良程度相似度计算方法。一张色谱指纹图谱可视为一个相似系统, 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。设两色谱指纹图谱相似要素组成的向量分别为 $(x_1, x_2, x_3, \dots, x_n)$ 和 $(y_1, y_2, y_3, \dots, y_n)$, 两图谱的改良程度相似度为 $Q_c = 1 - \frac{1}{n} \sum_{i=1}^n \left| 1 - \frac{x_i}{y_i} \right|$ 。式中 x_i 和 y_i 分别是两色谱指纹图谱对应峰峰面积, x_i/y_i 是相似元的相似程度。

2.2 改良程度相似度的改进

数学上用平均偏差或标准偏差表示一组数据的离散程度, 相对于平均偏差, 标准偏差可以突出较大偏差对测定结果重现性的影响。假设一组数值 $x_1, x_2, x_3, \dots, x_n$, 此组数值的标准差为:

200920622 收稿; 2009207231 接受
本文系北京市科技计划课题 (No D02050040040111) 资助
* E-mail: yjqia@263.net

$$R = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

(1)

平均偏差表示为:

$$d = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

(2)

改良程度相似度以对应峰面积比值参与计算, 对于不同批次的中药制剂, 两色谱图间对应峰的数目为 n , 对应峰面积或峰高为 x_i 和 y_i , 对应峰的比值 x_i/y_i 围绕 1 波动. 若将比值 x_i/y_i 视为一系列测量值, $1 - x_i/y_i$ 可以看成单个测量值的偏差, $\frac{1}{n} \sum_{i=1}^n \left| 1 - \frac{x_i}{y_i} \right|$ 相当于一系列测量值的平均偏差见式 (2). 标准偏差比平均偏差更能反映出数值中大偏差的影响, 参照式 (1) 将改良程度相似度 Q_c 改进为新改良程度相似度 q_c 见式 (3). 新改良程度相似度 q_c 适合对应峰峰面积差异相差不大的两图谱相似度的比较.

$$q_c = 1 - \sqrt{\frac{\sum_{i=1}^n \left(1 - \frac{x_i}{y_i} \right)^2}{n}}$$

(3)

当所有对应峰峰面积出现大差异 (即 $(1 - x_i/y_i)^2 \hat{=} 1$ 或 $x_i > 2y_i$) 时, q_c 出现负值而无意义. 若矢量中部分相似元 $x_i > 2y_i$, 相对差异越大的相似元对 q_c 影响越大. 当某个峰的相对差异超过十几倍或几十倍, $|1 - x_i/y_i|$ 为十几或几十, 该峰对 q_c 相似度的计算结果起主导作用, 相似度可能出现负值. 所以新改良程度相似度 q_c 适合对应峰峰面积差异相差不超过 100% 两色谱指纹图谱相似度的衡量.

2 3 数据来源

本研究以 1~19 号模拟样本和 9 个批次的乳块消片色谱指纹图谱数据为例比较改进后的相似度计算结果与原计算结果的差异. 乳块消片样本色谱指纹图谱数据的采集方法参见文献 [10], 相似度的计算程序自行编写.

3 结果与讨论

3 1 改进后的相似度与原相似度的比较

为了对改进后的相似度和原相似度计算结果进行比较, 表 1 采用差异明确的模拟数据来考察 4 种相似度的计算结果. 其中 10、16~19 号 5 个样本参考文献 [7] 的模拟数据, 与参照样本相比, 是各元素的相对差异完全相同的理想样本, 各元素的相对差异分别为 10%, 20%, , 50%. 实际情况中, 各元素的相对差异不完全相同, 相对差异大和相对差异小的元素共同存在, 因此本研究设计了 1~9 号和 11~15 号非理想的样本, 1~9 号样本和 11~15 号样本在参照样本的基础上部分元素出现 10%~60% 的差异, 产生差异的元素用黑体标示出.

表 1 模拟数据 4 种相似度的结果
Table 1 Results of four similarity algorithms based on simulated data

样本号 Sample number	样本的模拟数据 Simulated data of samples						夹角余弦 Cosine	相关系数 Correlation coefficient	改良程度相似度 Improved extent similarity	新改良程度相似度 q_c New improved extent similarity
参照 Reference	1	5	10	15	20	25				
1	1	5	9	15	22	25	0.9985	0.9955	0.9667	0.9423
2	1	5	10	15	24	25	0.9963	0.9969	0.9667	0.9184
3	1	5	9	16	5 22	25	0.9982	0.9931	0.9500	0.9293
4	1	5	9	15	24	25	0.9956	0.9941	0.9500	0.9087
5	1	5	10	15	26	25	0.9922	0.9973	0.9500	0.8775
6	0.9	5	9	16	5 22	25	0.9982	0.9851	0.9333	0.9184
7	0.9	5	9	15	24	25	0.9956	0.9941	0.9333	0.9000

续表 1(Continued to Table 1)

样本号 Sample number	样本的模拟数据 Simulated data of samples						夹角余弦 Cosine	相关系数 Correlation coefficient	改良程度相似度 Improved extent similarity	新改良程度 相似度 qc New improved extent similarity
8	0 9	5	10	15	26	25	0 9922	0 9973	0 9333	0 8709
9	1	5	10	15	28	25	0 9870	0 9981	0 9333	0 8367
10	1 1	4 5	11	13 5	22	22 5	0 9951	0 9837	0 9000	0 9000
11	1 2	4 5	11	13 5	22	25	0 9975	0 9920	0 9000	0 8845
12	1 3	4 5	11	13 5	20	25	0 9987	0 9958	0 9000	0 8586
13	1 4	4 5	11	15	20	25	0 9995	0 9985	0 9000	0 8268
14	1 5	4 5	10	15	20	25	0 9998	0 9994	0 9000	0 7918
15	1 6	5	10	15	20	25	0 9999	0 9998	0 9000	0 7551
16	1 2	4	12	12	24	20	0 9799	0 9334	0 8000	0 8000
17	1 3	3 5	13	10 5	26	17 5	0 9540	0 8543	0 7000	0 7000
18	1 4	3	14	9	28	15	0 9180	0 7585	0 6000	0 6000
19	1 5	2 5	15	7 5	30	12 5	0 8737	0 6592	0 5000	0 5000

黑体字表示产生了差异的数据 (The boldface numbers mean discrepant data)。

对表 1 中 19 个模拟样本进行分析, 按照样本矢量的特征可分为以下两大类: (1)各个元素的相对差异完全相同的理想样本, 如表 1 中 10、16~ 19 号样本, 夹角余弦和 pearson 相关系数相似度值远高于数值自身反映出的实际相似程度, 改良程度相似度和新改良程度相似度 qc 可以准确地反映样本与参照样本间的真实差异。(2)各个元素的相对差异不完全相同的样本, 如表 1 中 1~ 9 号样本和 11~ 15 号样本, 按照与参照样本的总差异的大小可以分为四组, 1~ 2 号样本、3~ 5 号样本、6~ 9 号样本和 11~ 15 号样本, 每组样本相对于参照样本的总差异相同, 差异分布不同, 各组样本的改良程度相似度值相同, 而新改良程度相似度 qc 能区分各组内差异分布不同的样本。

表 1 中 10 号样本各个元素变异 10%, 改良程度相似度和新改良程度相似度 qc 为 0 9000, 两种相似度的计算值与数值自身反映的差异一致。11~ 15 号样本与 10 号样本总差异相同, 均为 60%, 但差异的分布不同, 例如 11 号样本出现 1 个 20% 差异和 4 个 10% 差异, 12 号样本 1 个 30% 差异和 3 个 10% 差异, 13 号样本出现 3 个元素 20% 的差异, 而 15 号样本 1 个 60% 的大差异。11~ 15 号样本的改良程度相似度值不变, 但是随着大差异的出现, 相似度 qc 的计算值减小。

同样, 表 1 中 1~ 9 号样本是各个元素差异不完全相同的非理想样本, 1~ 2 号样本、3~ 5 号样本和 6~ 9 号样本相对于参照样本的总差异分别为 20%、30% 和 40%, 样本差异的分布不同。例如, 2 号样本有 1 个 20% 差异, 1 号样本有 2 个 10% 差异, 新改良程度相似度 qc 能突现 2 号样本 20% 差异, 2 号样本的相似度 qc 值比 1 号样本的 qc 值小。同样, 相对于参照样本, 3 号样本有 3 个 10% 差异的元素, 4 号样本具有 1 个 10% 差异和 1 个 20% 差异的元素, 5 号样本只有 1 个 30% 差异元素, 相似度 qc 能区分以上 3 个样本。若相似度以 0 9 为合格下限, 相似度 qc 能突出 5 号样本的 30% 大差异, 相似度值降至 0 18775 将 5 号样本判为不合格。与前两组样本类似, 6~ 9 号样本的差异由较均匀的分布逐渐集中到 1 个元素, 由 6 号样本 4 个 10% 的差异集中到 9 号样本仅 1 个 40% 大差异, 随着大差异的出现, 相似度 qc 值逐渐降低。

综上所述, 新改良程度相似度 qc 能突出样本中元素的大差异, 相似度 qc 的计算值予以降低, 这与中药质量控制的思路是一致的。中药所含成分按一定的配比存在时药效达到最优, 当各成分的量在既定配比合适的窗口内, 即在既定配比左右波动不大时, 对其药效的影响较小。当某个成分或某些成分偏离既定配比很远, 超出合适的配比窗口, 就会显著降低原成分群的药效。因此, 用相似度来衡量中药质量稳定性时应灵敏地反映组分群中某些组分对既定配比大的偏离, 新改良程度相似度正好符合这一要求, 适合药效成分群的配比波动的控制。

4 2 乳块消片色谱指纹图谱数据的相似度评价

采用 9 个批次的乳块消片指纹图谱中 28 个共有峰峰面积的中位数生成对照指纹图谱, 计算 9 批样品的色谱指纹图谱与对照指纹图谱的相似度, 结果表明, 改进后的 qc 相似度计算方法能够更精细地反映

样品间的相对差异。4种相似度计算结果见表 2。

由表 2可以看出,以夹角余弦和相关系数法计算的不同批次乳块消片色谱指纹图谱间的相似度都达到 0.998以上,没有明显的区别。改良程度相似度的计算值在 0.91~0.96之间,对 9批样品有了区分。改进后的相似度 q_d 的计算结果的区分度增大,相似度在 0.85~0.94之间,其中 73002、73005、73008和 73009与参照指纹图谱的改良程度相似度值在 0.93~0.94左右,而 73005和 73008批样品与参照指纹图谱的新改良程度相似度 q_d 值大于 0.9,73002和 73009批样品的相似度 q_d 值小于 0.9,说明相对于 73005和 73008批样品,73002和 73009批样品相对于参照指纹图谱某些峰存在大的差异。

表 2 9批乳块消样品与样品中位数对照色谱指纹图谱间的相似度
Table 2 Similarity of chromatographic fingerprints among nine batches Rukuaixiao samples and sample median

样品批号 Batch numbers of samples	相关系数 Correlation coefficient	夹角余弦 Cosine of angle	改良程度相似度 Improved extent similarity	新改良程度相似度 q_d New improved similarity
73002	1.000	1.000	0.9430	0.8933
73003	1.000	1.000	0.9397	0.9278
73004	0.999	0.999	0.9213	0.8524
73005	0.999	0.999	0.9391	0.9001
73006	1.000	1.000	0.9613	0.9403
73007	1.000	1.000	0.9523	0.9346
73008	0.999	0.999	0.9344	0.9103
73009	0.999	0.999	0.9407	0.8980
73010	0.998	0.999	0.9103	0.8766

4.3 新改良程度相似度 q_d 应用中需要注意的问题

改进后的新改良程度相似度 q_d 对应着其合适的使用范围。当共有峰峰面积差异不超过 100%,相似度 q_d 可以准确地反映样本与参照样本间的差异,适合不同批次中药制剂的色谱指纹图谱相似度的计算和中药生产过程的质量控制。对于中药材可能出现部分色谱峰的相对差异远大于 100%的情况,超出了相似度 q_d 的适用范围,因此在对药材的指纹图谱进行相似度计算时需采用其它方法,这与文献[7]对改良程度相似度的适用性分析是一致的。

综上所述,在共有峰峰面积差异不超过 100%的情况下,改进后的新改良程度相似度 q_d 能突现样本里的大差异,区分与参照样本的总差异相同差异分布不同的样本,灵敏地反映成分群中某些成分对既定配比大的偏离,适用于对成分群配比波动的控制。本研究所建立的相似度计算方法能反映中药内在质量的稳定性,可基于化学物质基础进行中药生产过程的质量控制。

References

1 SUN GuoXiang(孙国祥), HOU Zhizhi(侯志飞), ZHANG Chunling(张春玲), BIKaizhun(毕开顺), SUN YuQin(孙毓庆). Acta Pharmaceutica Sinica(药学学报), 2007, 42(1): 75~80

2 HE Fuyuan(贺福元), ZHOU Honghao(周宏灏), DENG Kaiwen(邓凯文), LIU Wenlong(刘文龙), LIU Ping'an(刘平安). Acta Pharmaceutica Sinica(药学学报), 2008, 43(2): 195~201

3 HUANG Sheng(黄胜), HE Fuyuan(贺福元), LIU Wenlong(刘文龙), LI Qiquan(李启泉), JIANG Huanying(江欢英), NIE Zhizhao(聂志姣), LUO Jieying(罗杰英). Chinese Traditional Patent Medicine(中成药), 2009, 31(1): 35~39

4 Xiang SM, Nie FP, Zhang CS. Pattern Recognition, 2008, 41(12): 3600~3612

5 Wang K, Du K, Li H. Computers and Applied Chemistry, 2007, 24(1): 49~52

6 LIU YongSuo(刘永锁), MENG Qinhua(孟庆华), JIANG Shumin(蒋淑敏), HU Yuzhu(胡育筑). Chinese Journal of Chromatography(色谱), 2005, 23(2): 158~163

7 LIU YongSuo(刘永锁), CAO Min(曹敏), WANG Yiming(王义明), LUO Guoan(罗国安). Chinese J. Anal. Chem. (分析化学), 2006, 34(3): 333~337

8 ZHOU Meili(周美立), WANG Huanchen(王浣尘). Systems Engineering(系统工程), 1996, 14(4): 1~6

- 9 ZHOU MeL i(周美立), WU BaoRen(吴报任). Journal of Anhui Institute of Technology (安徽工学院学报), 1991, 10(3): 67~ 73
- 10 ZHAN X iaoR i(展晓日), SH I X i n2Yuan(史新元), ZHANG Pei(张培), Q IAO Y an2Jiang(乔延江). Chinese Journal of Experimental Traditional Medical Formulae(中国实验方剂学杂志), 2008, 14(4): 1~ 3

Improvements in Similarity Algorithms Based on Similarity System Theory

ZHAN Xue2Yan, SH I X i n2Yuan, ZHAN X iaoR i, WANG Yun, Q IAO Y an2Jiang*

(Pharmacy College of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100102)

Abstract For solving the problem that common similarity algorithms are insensitive to the relative differences of different chromatographic fingerprints data, this paper improved the algorithms of improved extent similarity, which named new improved extent similarity. It was confirmed that new improved extent similarity could distinguish the samples with the same total difference and different distribution of relative differences by applying the similarity algorithm to simulated data and the chromatographic fingerprints of nine batches Ruku2aixiao tablets. New improved extent similarity could reflect the relative differences of peak areas more sensitively when the relative differences of common peak areas was less than 100%. The result indicates that new improved extent similarity can reflect the large deviation from the established proportion of components, which can be applied to the control of the fluctuation of the established proportion of effective components and the process quality control of Traditional Chinese Medicine.

Keywords Similarity system theory; Similarity algorithms; Process quality control

(Received 22 June 2009; accepted 31 July 2009)

全国生物医药色谱学术交流会 (2010)征文

由中国化学会色谱专业委员会和北京理化分析测试技术学会北京色谱学会主办的/全国生物医药色谱学术交流会(2010)定于2010年5月7~11日在景德镇市召开。

会议将就色谱技术在生命科学、生物技术、药物、环境、食品安全等领域的分离、分析和质量控制等方面的发展和应

用进行学术研讨,热诚欢迎全国从事生物医药色谱工作的科技人员及相关企业与会交流。

论文摘要请于2010年3月31日前向本会网站投稿。会议的部分征文将以全文在5化学通报6发表,请有意发表全文的作者,除向网站投递摘要外,按照5化学通报6的格式准备好全文文稿,于2010年3月31日前用电子邮件发送到会议学术组。

欢迎国内外分析仪器公司和厂商到会介绍和展出产品,具体事宜请与会务组联系。

会议网站: <http://www.cbm.cn.org>(2010年二月初启用)

学术组联系人:赵睿,010-62557910 Email: zhaou@iccas.ac.cn 中国科学院化学研究所,北京中关村北一街2号,邮编100190

会务组联系人:于靖琦,010-68731259,桂三刚,010-88417672,传真:010-68471169, Email: VIP001@21cn.com;北京理化分析测试技术学会,北京海淀区西三环北路27号,邮编100089