Vol. 27 高等学校化学学报 No. 1
2006 年 1 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 26~29

基于信息融合的中药多元色谱指纹图谱相似性计算方法

范骁辉, 叶正良, 程翼宇
(浙江大学药学院中药科学与工程学系, 药物信息学研究所, 杭州 310027)

摘要 用信息融合算法合并多张色谱指纹图谱, 解决中药多元指纹图谱相似性评价难题, 提出一种多元色谱指纹图谱相似度计算方法. 用该方法先对各单元指纹图谱进行串行或并行象素级信息融合, 再对融合了各指纹峰信息的多元色谱指纹图谱进行相似度评价. 计算机仿真和复方丹参滴丸多元 HPLC 指纹图谱应用结果表明, 该法能够评价中药多元色谱指纹图谱相似度, 定量表征中成药产品批次间质量波动情况.

关键词 多元化学指纹图谱; 信息融合; 中药质量控制
中图分类号 O657.7; R284 **文献标识码** A **文章编号** 0251-0790(2006)01-0026-04

化学指纹图谱分析技术[1~4]是科学可行的现代中药质量控制方法, 已引起分析化学界的广泛关注, 并为国家药监部门所重视. 目前所见的中药化学指纹图谱大多是单张的色谱指纹图谱. 众所周知, 当药品化学物质体系比较复杂时, 单张化学指纹图谱往往难以完整地反映出中药产品的化学组成特征, 需要建立多元化学指纹图谱[3,4]. 即将多张反映药品若干部分化学组成特征的指纹图谱组合在一起, 共同表征药品完整的化学组成特征, 形成多元指纹图谱. 现行的中药指纹图谱相似度计算方法[5]适宜于待测指纹图谱与标准指纹图谱间相似性的一对一定量评价, 而不适用于指纹图谱相似性的多对多比较, 故难以用于定量评价这类多元指纹图谱的相似性.

本文提出了一种基于信息融合的中药多元色谱指纹图谱相似性的计算方法. 计算机仿真和中药产品实际应用结果表明, 该方法能够评价多元色谱指纹图谱相似度, 定量表征中成药产品批次间质量波动情况.

1 基本原理

1.1 多元色谱指纹图谱的信息融合法

信息融合[6]是对人脑综合处理信息过程的模拟, 已被广泛应用于医学图像处理和生物传感器信息处理等领域. 其基本思想是根据某种准则对多元的观测信息进行整合, 以获得被测对象的合理解释或描述. 象素级融合是目前最常见的信息融合方法, 它通过对原始数据直接进行组合得到融合向量. 其中, 向量的组合策略又分成串行和并行两种.

以两张指纹图谱组成的多元化学指纹图谱为例, 其串行和并行象素级信息融合方法分别如下: 以向量 $\boldsymbol{\alpha}_1=[\alpha_{11},\alpha_{12},\cdots,\alpha_{1m}]$ 和 $\boldsymbol{\alpha}_2=[\alpha_{21},\alpha_{22},\cdots,\alpha_{2n}]$ 分别表示多元色谱指纹图谱的指纹图谱 Ⅰ 和指纹图谱 Ⅱ (其中, α_{1i} 和 α_{2i} 为采样数据值或谱峰面积值). 当采用串行方式对多元化学指纹图谱进行象素级融合时, 融合后的向量 $\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta\boldsymbol{\alpha}_2]$; 而并行融合结果 $\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta\boldsymbol{\alpha}_2]$ (当向量 $\boldsymbol{\alpha}_1$ 与 $\boldsymbol{\alpha}_2$ 维数不等时, 低维向量一般通过增加零值来补足维数), 其中, θ 为组合系数, 其值一般视具体情况而定.

如果多元色谱指纹图谱由 3 张谱图组成, 其串行融合结果 $\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta_1\boldsymbol{\alpha}_2,\theta_2\boldsymbol{\alpha}_3]$, 并行融合结果

收稿日期: 2005-04-29.
基金项目: 国家自然科学基金重大研究计划重点项目(批准号: 90209005)和国家"十五"重大科技攻关计划项目(批准号: 2001BA701A01)资助.
联系人简介: 程翼宇(1958 年出生), 男, 博士, 教授, 博士生导师, 从事药物分析和药物信息学研究.
E-mail: chengyy@zju.edu.cn

$\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta_1\boldsymbol{\alpha}_2,\theta_2\boldsymbol{\alpha}_3]$；依此类推，由 n 张谱图组成的多元色谱指纹图谱，其串行融合向量 $\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta_1\boldsymbol{\alpha}_2,\cdots,\theta_{n-1}\boldsymbol{\alpha}_n]$，其并行融合结果 $\boldsymbol{\phi}=[\boldsymbol{\alpha}_1,\theta_1\boldsymbol{\alpha}_2,\cdots,\theta_{n-1}\boldsymbol{\alpha}_n]$.

此外，由于样品预处理方法以及分析仪器等的不同，各谱图向量在绝对数值上可能存在较大的差异. 因此，一般还需要先对原始向量分别进行归一化处理：

$$\boldsymbol{\alpha}'_i=\boldsymbol{\alpha}_i\Big/\sum_{i=1}^{n}\boldsymbol{\alpha}_i \tag{1}$$

式中，$\boldsymbol{\alpha}'_i$是归一化后的峰面积值或采样数据值，$\boldsymbol{\alpha}_i$是某一子指纹图谱的第 i 个峰的面积值或采样数据值.

1.2　多元色谱指纹图谱相似度计算法

上述化学信息融合结果反映了多元色谱指纹图谱的整体信息，因此只需根据得到的融合向量计算多元色谱指纹图谱相似度，即可对药品生产批次间质量稳定性做出评价. 文中采用夹角余弦测度[7]来计算各产品多元色谱指纹图谱间的相似度.

$$S(\boldsymbol{\phi}_i,\boldsymbol{\phi}_0)=\Big(\sum_{k=1}^{m}\boldsymbol{\phi}_{ik}\cdot\boldsymbol{\phi}_{0k}\Big)\Big/\Big(\sum_{k=1}^{m}\boldsymbol{\phi}_{ik}^2\sum_{k=1}^{m}\boldsymbol{\phi}_{0k}^2\Big) \tag{2}$$

式中，$\boldsymbol{\phi}_i$为待测样品多元色谱指纹图谱的信息融合结果；$\boldsymbol{\phi}_0$为对照用多元色谱指纹图谱信息融合结果.

2　计算机仿真研究

为证明技术的可行性，采用计算机仿真手段考察多元色谱指纹图谱象素级信息融合结果. 仿真指纹图谱采用高斯模型产生，数学模型表达式为

$$\boldsymbol{\alpha}_i(t)=\sum_{i=1}^{N}\frac{A_i}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t-t_i)^2}{2\sigma^2}\right) \tag{3}$$

式中，N 为总峰数，t_i为第 i 峰保留时间，A_i为第 i 峰面积值，σ 为高斯函数标准偏差.

为简单起见，假定多元色谱指纹图谱由两张谱图组成. 设指纹图谱Ⅰ为 $\boldsymbol{\alpha}_1=[1,3]$；指纹图谱Ⅱ为 $\boldsymbol{\alpha}_2=[3,1]$. 将参数分别代入式(3)模拟产生两张谱图组成多元色谱指纹图谱. 若 θ_1取 1，信息融合结果如图 1 所示，其串行融合结果 $\boldsymbol{\phi}=[3,1,1,3]$，并行融合结果 $\boldsymbol{\phi}=\begin{vmatrix}3&1\\1&3\end{vmatrix}$.

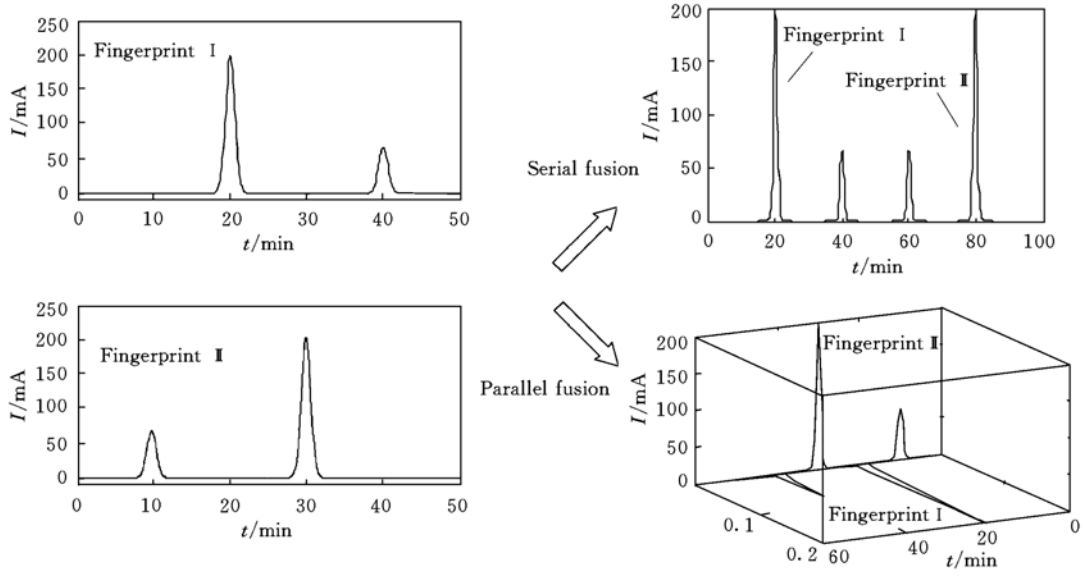

Fig. 1　Illustration of serial and parallel fusion for multiple chromatographic fingerprints

分别改变指纹图谱Ⅰ和Ⅱ中各峰面积大小，按式(2)计算各单元指纹图谱及多元色谱指纹图谱相似度值. 表 1 给出了部分仿真结果，可见，基于两种融合策略的多元色谱指纹图谱相似度结果一致，

均能体现各单元指纹图谱的变化情况. 以此类推, 如果多元色谱指纹图谱由 2 张以上谱图组成, 同样可以得到类似的结果.

Table 1 The similarity of simulated multiple chromatographic fingerprints

No.	Fingerprint Ⅰ	Fingerprint Ⅱ	Serial fusion	Parallel fusion
1	1.0	0.6	0.8	0.8
2	0.6	1.0	0.8	0.8
3	0.6	0.6	0.6	0.6

3 实例研究

为验证本方法的实用性, 选择复方丹参滴丸 HPLC 指纹图谱为实例进行考察. 复方丹参滴丸样品由天津天士力公司提供. 其中, 制药工艺参数正常产品 13 批, 编号 1 ~ 13 ; 工艺参数异常产品 7 批, 编号 14 ~ 20. 研究[7]表明, 复方丹参滴丸的主要药效成分为丹参水溶性成分和三七皂苷类成分. 在获取 HPLC 指纹图谱时发现, 难以选定合适的分离条件得到一张可同时完整反映这两类成分的 HPLC 谱图, 故需分别建立 HPLC 指纹图谱(图 2), 从而用多元 HPLC 指纹图谱来表征产品的化学组成. 这显然涉及到多元色谱指纹图谱相似度计算问题, 因此该实例具有典型性.

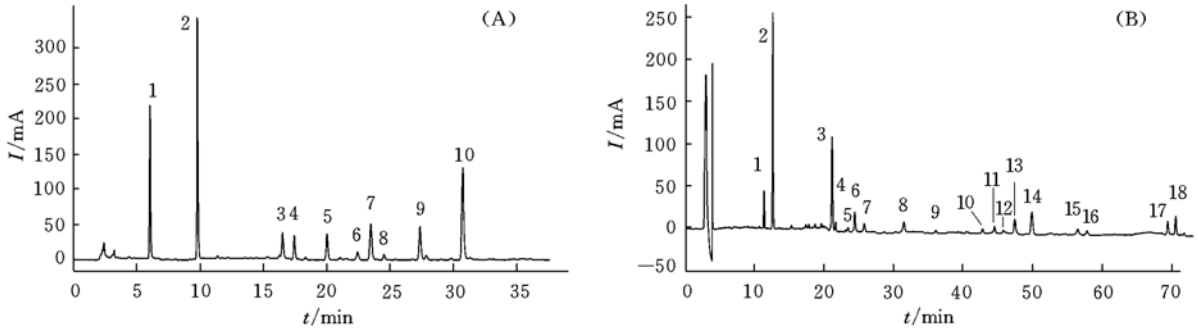

Fig. 2 Multiple HPLC fingerprints of CDDP

(A) Fingerprint Ⅰ; (B) Fingerprint Ⅱ.

3.1 组合系数 θ 的确定

在医学图像处理和人脸识别等信息融合技术的传统应用领域中, 组合系数 θ 通常取为 1 或各组特征维数之比[8]. 中药多元色谱指纹图谱的 θ 值最好根据药效成分指纹峰情况来确定. 本文所述 HPLC 多元指纹图谱分别表征复方丹参滴丸的两类药效成分: 指纹图谱Ⅰ反映的是复方丹参滴丸中酚酸类成分的化学组成特征, 指纹图谱Ⅱ反映了复方丹参滴丸中三七皂苷类成分的化学组成特征. 显然, 若取 θ 为 1, 由于含量差异较大, 将两类成分完全等同视之并不合理. 本文取 θ 为指纹图谱Ⅱ所代表的化学组分与指纹图谱Ⅰ所代表的化学组分在样品中含量的比值. 进一步的化学物质基础研究结果表明[9], 指纹图谱Ⅰ表征的丹参酚酸类成分约占复方丹参滴丸的 5%; 指纹图谱Ⅱ表征的三七皂苷类成分约占复方丹参滴丸的 11% . 据此, 本文暂取 θ 为 2.2. 事实上, 如何精确估计组合系数, 使得组合后的效果最佳仍是一个有待研究的问题.

3.2 基于象素级信息融合的多元色谱指纹图谱相似度计算

分别测取复方丹参滴丸 HPLC 指纹图谱Ⅰ中 10 个和指纹图谱Ⅱ中 18 个共有峰的面积值用于计算. 按式(2)归一化后, 分别采用串行和并行方式进行信息融合, 并以制药工艺参数正常产品(1 ~ 13 号样品)指纹图谱的均值作为标准指纹图谱, 按式(2)计算相似度值, 结果如图 3 所示.

由图 3 可见, 基于并行象素级融合的多元色谱指纹图谱相似度计算方法中通过增加零值来补足维数, 其相似度计算结果稍低于基于串行象素级信息融合的相似度值, 但两种方法的计算结果趋势基本一致: 工艺参数正常产品的相似度值较高(>0.9), 而工艺参数异常产品的相似度值较低(<0.8). 这表明, 基于串行象素级信息融合和基于并行象素级融合的多元色谱指纹图谱相似度计算方法均能反映复方丹参滴丸生产批次间的质量差异, 从而可区分出制药工艺参数正常和异常产品.

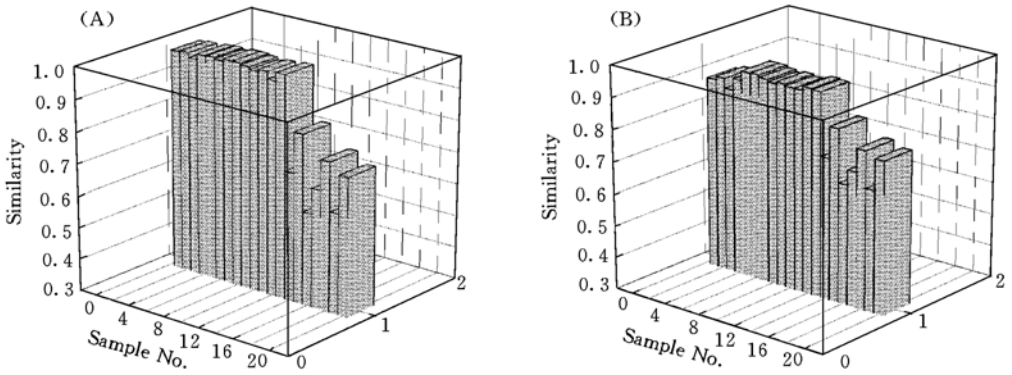

Fig. 3 Similarity of multiple HPLC fingerprints of CDDP

(A) Serial information fusion; (B) parallel information fusion.

参 考 文 献

[1] CHEN Min-Jun(陈闽军), QU Hai-Bin(瞿海斌), CHENG Yi-Yu(程翼宇). Chem. J. Chinese Universities(高等学校化学学报)[J], 2003, **24**(12): 2181—2185

[2] World Health Organization. General Guidelines for Methodologies on Research and Evaluation of Traditional Medicines[EB/OL], Geneva, 2000

[3] Food and Drug Administration. Guidance for Industry Botanical Drug Products(Draft Guidance)[EB/OL], Rockville, 2000

[4] State Food and Drug Administration(国家药品食品监督管理局). Drug Standards of China(中国药品标准)[J], 2000, **1**(4): 3—7

[5] CHENG Yi-Yu(程翼宇), CHEN Min-Jun(陈闽军), WU Yong-Jiang(吴永江). Acta Chim. Sinica(化学学报)[J], 2002, **60**(11): 2017—2021

[6] Willett P., Holliday J., Salim N.. J. Chem. Inf. Comput. Sci.[J], 2003, **43**: 435—442

[7] MA Jian-Wen(马建文), YE Zheng-Liang(叶正良). World Science and Technology: Modernization of TCM(世界科学技术：中药现代化)[J], 2000, **2**(2): 44—47

[8] YANG Jian(杨 建), YANG Jing-Yu(杨靖宇), GAO Jian-Zhen(高建贞). Journal of Software(软件)[J], 2003, **14**(3): 490—495

[9] ZHANG Hai-Jiang(张海江). Doctor Degree Dissertation[D], Zhejiang University, 2004

A Computational Method Based on Information Fusion for Evaluating the Similarity of Multiple Chromatographic Fingerprints of TCM

FAN Xiao-Hui, YE Zheng-Liang, CHENG Yi-Yu*

(*Pharmaceutical Informatics Institute, Department of Chinese Medicine Science and Engineering, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou* 310027, *China*)

Abstract An computational method based on information fusion is developed to evaluate the similarity of multiple chromatographic fingerprints of TCM, where it solves a current key problem on evaluating the similarity of multiple fingerprints. In this method, the information of individual fingerprint was combined by using serial or parallel information fusion strategy at raw data level, and then the integral similarity of multiple fingerprints was calculated by comparing their fusion results. Subsequently, this method was applied to simulated datasets and a set of Compound Danshen Dripped Pills(CDDP) samples. The results indicate that, by this method, the similarity among multiple chromatographic fingerprints of TCM can be calculated, and the lot-to-lot consistency of samples can be evaluated quantitatively.

Keywords Multiple chromatographic fingerprinting; Information fusion; Quality control of TCM

(Ed.: K, G)