·论 著·

中药色谱指纹图谱全排序模板匹配算法研究

朱臻宇,乔善磊,张 海,娄子洋,柴逸峰*
(第二军医大学药学院药物分析学教研室,上海 200433)

[摘要] **目的**:提出一种全排序模板匹配算法,用于中药色谱指纹图谱的色谱峰匹配和指纹图谱相似度计算。**方法**:以高效液相色谱法(HPLC)测定丹参药材中丹参酮ⅡA为例,分别采集样品溶液的色谱图,对色谱峰进行匹配,在原有算法基础上,提出一种全排序模板匹配算法,这种算法不以任何一张实际的指纹图谱为模板,而是根据所有图谱上的所有色谱峰的保留时间进行升序排序、聚类,依从色谱峰最接近、同一组匹配峰中不包含两个同源峰、待匹配的峰不超过匹配阈值等原则。**结果**:这种方法克服了原算法中的某些错配、漏配等缺陷,得到了较满意的匹配结果。**结论**:此算法为以保留时间为依据的色谱峰匹配算法找到了新的切入点。

[关键词] 全排序模板匹配算法;色谱指纹图谱;中草药
[中图分类号] R 28 [文献标识码] A [文章编号] 0258-879X(2007)02-0183-05

A modified template matching algorithm for chromatographic fingerprint of *S. miltiorrhiza* Bge, a traditional Chinese herb

ZHU Zhen-yu, QIAO Shan-lei, ZHANG Hai, LOU Zi-yang, CHAI Yi-feng* (Department of Pharmaceutical Analysis, School of Pharmacy, Second Military Medical University, Shanghai 200433, China)

[**ABSTRACT**] **Objective**: To introduce a new template peak matching algorithm for calculating the similarity of chromatographic fingerprint of traditional Chinese herbs. **Methods**: Tanshinone ⅡA in *S. miltiorrhiza* Bge of different batches were determined by HPLC and the chromatograms of them were obtained. We designated a new peak matching algorithm based on the previous algorithms, which employed a certain real chromatographic fingerprint as their template. In the new algorithm, we arranged the retention times of chromatographic peaks of all chromatograms into an ascending order, forming a template. The sorting procedure complied with the following 2 rules. First, the same peak matches the nearest corresponding peak. Second, one peak does not appear twice in the same chromatogram. **Results**: The new algorithm avoided the shortcomings of previous algorithm, such as mismatching and missed matching, and obtained a satisfactory outcome. **Conclusion**: Our new algorithm provides a basis for improving the reliability of retention time-based peak matching algorithm.

[**KEY WORDS**] total sequencing template matching algorithm; chromatographic fingerprint; drugs, Chinese herbal
[Acad J Sec Mil Med Univ,2007,28(2):183-187]

中药色谱指纹图谱技术是当前进行复杂物质体系质量控制及研究的重要手段,其对中药质量稳定性评价的方法已被国际社会所接受[1-3]。采用计算机辅助法对中药色谱指纹图谱进行相似度计算,关键环节之一是进行色谱峰的匹配。目前国内《中药注射剂色谱指纹图谱实验研究技术指南(试行)》[4]推荐使用色谱保留时间作为定性依据,多采用模板匹配算法。如直接采用时间窗宽度匹配[5-7],比较简单、思路清晰,但对模板上没有的峰不做匹配,导致匹配模板上没有的峰最终不参加相似度计算;时间窗法结合峰高信息的色谱峰匹配算法[8]对此做了改进,但如果待鉴定峰的附近有较多的峰高(峰面积)相接近的峰,仍可能发生峰次序匹配颠倒的问题。

上述两种算法都使用模板匹配的方法进行所有指纹图谱的峰匹配,而本文提出的全排序模板匹配算法不使用任何一张实际的指纹图谱作为模板,而是根据所有图谱上的所有峰的保留时间进行升序排序、聚类,克服了原算法中某些错配、漏配等缺陷,得到了较为满意的匹配结果。

1 方法原理

以保留时间作为定性依据,则保留时间一致的色谱峰应当被认为是同一物质的色谱峰,因此,当对所有待匹配的色谱峰依据保留时间排序后进行匹配时,同一物质的色谱峰应当彼此靠近(条件一);同

[基金项目] 上海市科委重点项目(03DZ19548);上海市科技发展基金(02419125). Supported by Fund of Key Program of Science Committee of Shanghai Municipal Government(03DZ19548) and Foundation for Science and Technology Development of Shanghai (02419125).

[作者简介] 朱臻宇,博士,讲师.
* Corresponding author. E-mail:yfchai@smmu.edu.cn

时,在一组匹配峰中,同一指纹图谱不应出现两个同源峰(条件二)。基于此思路,提出全排序模板匹配算法,其逻辑过程如图 1。

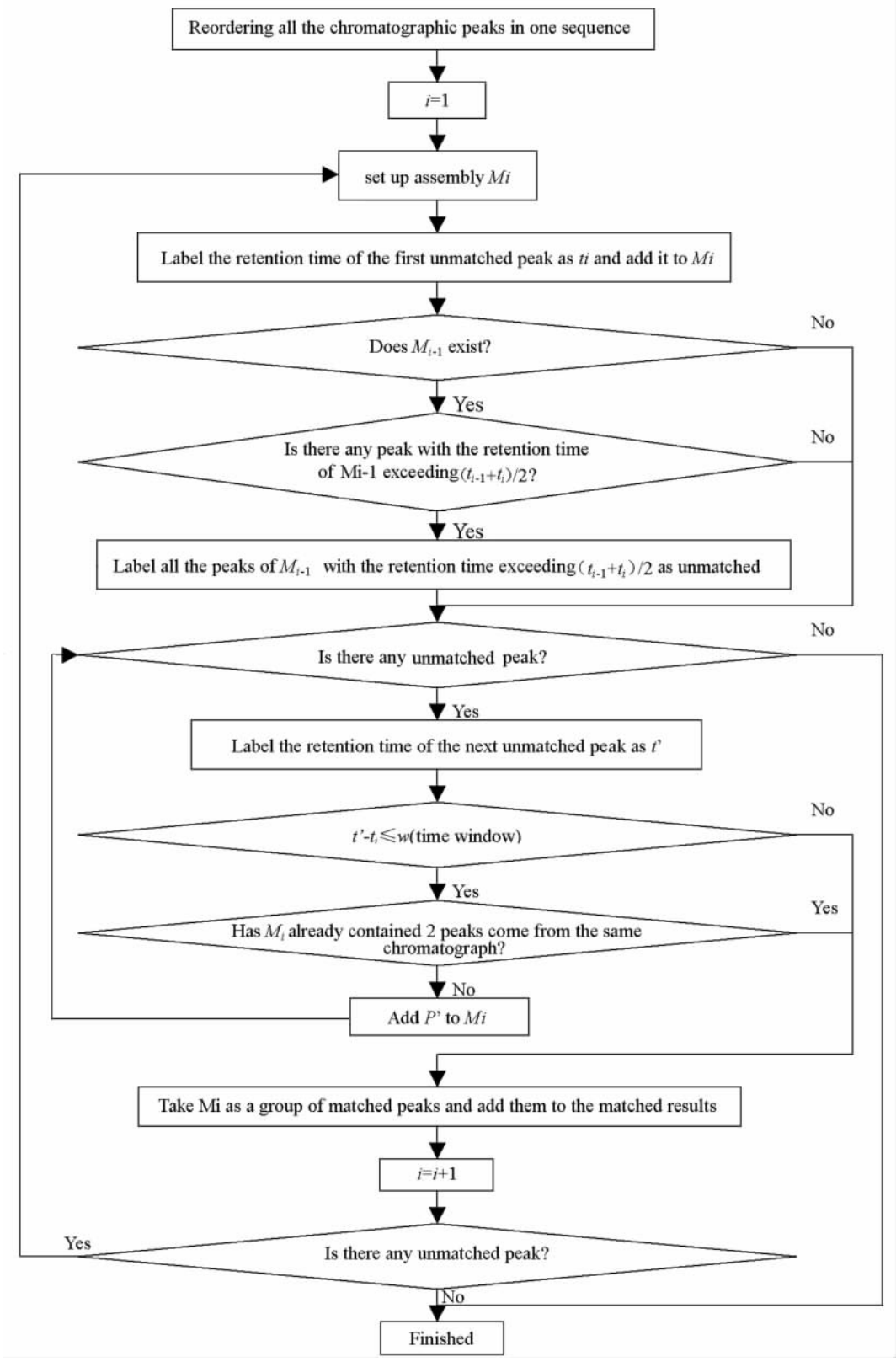

图 1　全排序模板匹配算法流程图

Fig 1　Flow chart of template matching algorithm

以下面一组数据为例说明全排序模板匹配算法的运算过程。3 张图各有 2 个色谱峰,保留时间分别为 3.09 min 和 3.15 min;3.05 min 和 3.10 min;3.10 min 和 3.15 min。时间窗 $w=0.2$ min。

运算步骤如下:1)全排序:3.05(2), 3.09(1), 3.10(2), 3.10(3), 3.15(1), 3.15(3);2)建立 M_1 集合,加入 3.05(2),令 $t_1=3.05$;3)开始依次向 M_1 中加入其他峰;4)3.09(1),满足条件;5)3.10(2),不满足条件二,M_1 加入完毕,记录 M_1[3.05(2),3.09(1)];6)建立 M_2 集合,加入 3.10(3),令 $t_2=3.10$;7)从 M_1 中删除保留时间 $>(t_1+t_2)/2$ 的峰 3.09(1);8)重建 M_2 集合,加入 3.09(1),令 $t_2=3.09$;9)开始依次向 M_2 中加入其他峰;10)3.10(2),满足条件;11)3.10(3),满足条件;12)3.15(1),不满足条件二,M_2 加入完毕,记录 M_2[3.09(1),3.10(2),3.10(3)];13)建立 M_3 集合,加入 3.15(1),令 $t_3=3.15$;14)M_2 中没有需要删除的峰,开始向 M_3 中加入其他峰;15)3.15(3),满足条件。

匹配结果:M_1[3.05(2)],M_2[3.09(1),3.10(2),3.10(3)],M_3[3.15(1),3.15(3)]。

2 材料和方法

2.1 *仪器和试剂* Waters 高效液相色谱仪,515 泵,996 二极管阵列检测器,Millennium^{32} 软件(美国 Waters 公司)。甲醇(色谱纯),石油醚(分析纯),重蒸水。丹参酮ⅡA 对照品(供含量测定用,中国药品生物制品检定所,批号 0766-200011)。样品:1~3 号丹参(S. miltiorrhiza Bge.)药材分别采自辽宁凌源、河北行塘、山东平邑(均采于 1999 年),由本院生药学教研室陈万生副教授鉴定。图谱保留时间校正和相似度计算使用第二军医大学和清华大学联合开发的《中药指纹图谱工作站》软件进行。

2.2 *色谱条件* 色谱柱:Hypersil C_{18} 柱(4.6 mm×200 mm,5 μm),大连依利特分析仪器有限公司制造;流动相:乙腈-25 mmol/L 磷酸二氢钠(磷酸调 pH 至 2.5),梯度洗脱程序为 0~34 min,乙腈的体积分数由 8%线性增至 30%;34~74 min,乙腈由 30%线性增至 78%;74~90 min,乙腈保持 78%。流速 1.0 ml/min;检测波长:280 nm,柱温:30℃。

2.3 *溶液的配制* 取丹参酮ⅡA 对照品 5 mg,精密称定,置 10 ml 量瓶中,加少量氯仿溶解,再加甲醇稀释至刻度,摇匀;精密量取 2.0 ml,置 10 ml 量瓶中,加甲醇至刻度,摇匀,即得每 1 ml 中含丹参酮ⅡA 100 μg 的对照品溶液。取丹参粉末约 1 g,精密称定,置索氏提取器中,加入石油醚(60~90℃)150 ml,置水浴上加热回流 6 h,放冷,滤过,滤液置 250 ml 烧瓶中,用石油醚洗涤残渣与滤纸,洗液滤入同一烧瓶中,将滤液减压浓缩至近干,以甲醇溶解转移至 50 ml 量瓶中,加甲醇稀释至刻度,摇匀即得供试品溶液。

2.4 *样品测定* 分别精密量取对照品溶液及供试品溶液(进样前将溶液过 0.45 μm 微孔滤膜)各 20 μl,注入高效液相色谱仪,使用 2.2 项下的色谱条件分析。

3 实验结果

色谱峰匹配是中药指纹图谱相似度软件实现的关键步骤,是得到可靠的相似度计算结果的保证。本文截取丹参色谱图 18~28 min 的一段,依从下列 3 个准则:(1)如果只有 1 个样品峰落在某个已知化合物的匹配时间窗内,该峰被认为是该化合物的峰;(2)如果有 1 个以上的样品峰落在某已知化合物的匹配时间窗内,与已知化合物保留时间最接近的峰被认为是该化合物的峰(即"色谱峰最接近"原则,也即"同一组匹配峰中不包含两个同源峰"原则);(3)如果待匹配的峰不落在任何待匹配化合物的时间窗内,该峰被列为未知化合物的峰(即"不超过匹配阈值"原则),采用全排序模板匹配算法对表 1 的 Ext_5、Ext_6、Ext_12 等三个色谱图进行全排序,得到所有保留时间的全排序模板匹配结果,如图 2。

表 1 色谱数据和匹配结果

Tab 1 Chromatographic data and their matching results

Chromatography	Peak No.						
	1	2	3	4	5	6	7
Ext_5	18.41	20.80	24.19	24.86	25.77	27.13	27.77
Ext_6	18.28	20.76	24.18	24.85	25.74	27.10	27.75
Ext_12	-	20.68	24.12	24.78	25.64	27.06	27.72

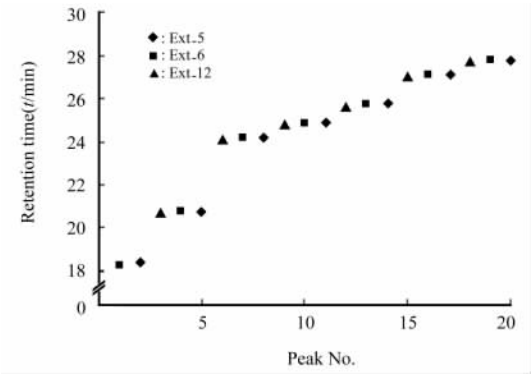

图 2　所有保留时间的全排序模板匹配结果

Fig 2　Matching results of all retention times

时间窗宽度匹配算法、时间窗法结合峰高信息的色谱峰匹配算法都需要从待匹配图谱中选择一条作为初始的匹配模板，因此统称之为模板匹配算法。模板上的色谱峰列表被当作“已知化合物”列表，取其它的指纹图谱使用时间窗法与之匹配，在待匹配指纹图谱中寻找与“已知化合物”相对应的色谱峰。由于它们使用了“从前至后依次匹配”的方式，如果匹配时间窗过大，将出现匹配错误。如图 3 所示，时间窗宽度匹配算法（w＝2.0 min）出现了明显的不合理的匹配结果，而使用全排序模板匹配算法则不会导致这种错误。如图 4。

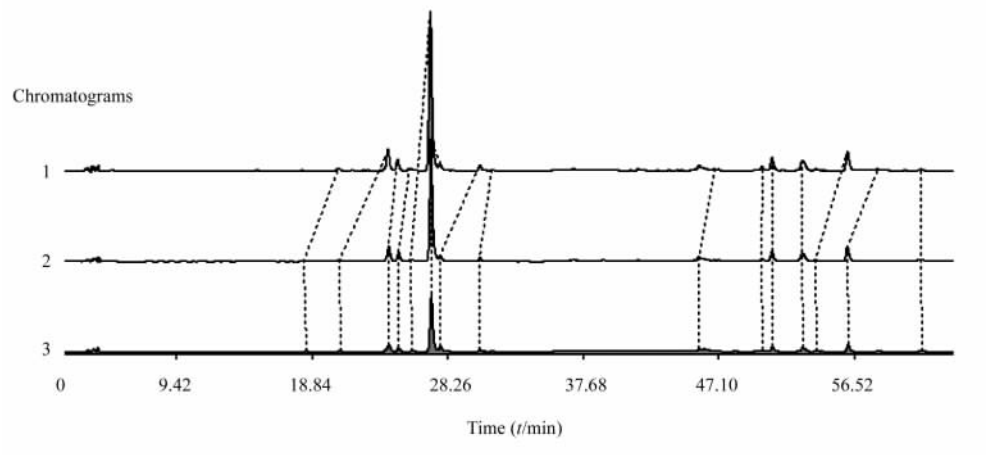

图 3　采用时间窗匹配算法的匹配结果——峰过宽（w＝2.0 min）引起匹配错误

Fig 3　Matching result of time windows width matching algorithm—matching errors caused by excessively wide peak (w＝2.0 min)

1：Fingerprint of *Danshen* from Liaoning Lingyuan；2：Hebei Xingtang；3：Shandong Pingyi

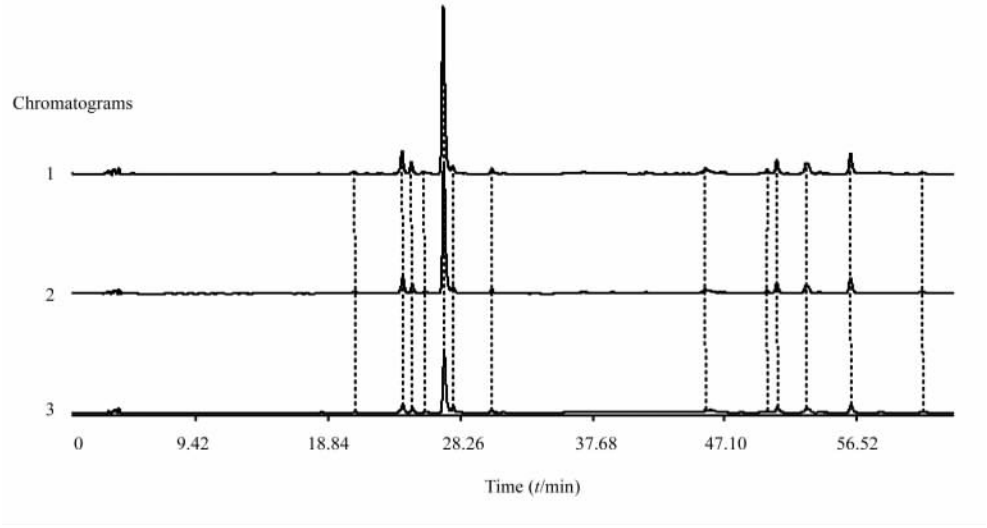

图 4　采用全排序模板匹配算法的匹配结果——即使峰宽达到 2.0 min 也不出现匹配错误

Fig 4　Matching result of template matching algorithm—no matching errors even if the peak is excessively wide (w＝2.0 min)

1：Fingerprint of *Danshen* from Liaoning Lingyuan；2：Hebei Xingtang；3：Shandong Pingyi

4 讨 论

本文提到的 3 种算法都只使用了保留时间作为色谱峰的定性依据,如果进行足够的算法修正和微小的判断准则的统一,则几种算法有可能得到统一的结果。由于保留时间本身的不可靠性,造成了这些算法都存在一定的不可靠性。更完善的算法应当是在现有依据保留时间的匹配方法的基础上,最大限度地利用更全面可靠的光谱信息,提出更合理的匹配算法。

[参 考 文 献]

[1] 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: 40-54.

[2] 李永国,王峥涛. 中药色谱指纹图谱的数据处理与指标[J]. 中药新药与临床药理,2005,16: 15-19.

[3] 郑 红,杜建卫. 中药指纹图谱的识别[J]. 北京石油化工学院学报, 2005,13: 56-60.

[4] 国家药品监督管理局. 关于印发《中药注射剂指纹图谱研究的技术要求(暂行)》的通知[J]. 中成药,2000,22:671-675.

[5] Pino J A, McMurry J E, Jurs P C,et al. Application of pyrolysis/gas chromatography/pattern recognition to the detection of cystic. Fibrosis Heterozygotes[J]. Anal Chem, 1985, 57: 295-302.

[6] Parrish M E, Good B W, Hsu F S, et al. Computer-enhanced high-resolution gas chromatography for the discriminative analysis of tobacco smoke [J]. Anal Chem, 1981, 53: 826-831.

[7] Johnson K J, Wright B W, Jarman K H, et al. A high-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis[J]. J Chromatogr A,2003,996: 141-155.

[8] 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: 713-717.

[收稿日期] 2006-07-10 [修回日期] 2006-12-27

[本文编辑] 尹 茶

· 读者 作者 编者 ·

作者标注中图分类号须知

为了便于检索和编制索引,本刊按《中国图书资料分类法》(4 版)标注论文的[中图分类号]。[中图分类号]置于中文[关键词]下方,单独起行。

《中国图书资料分类法》分为马克思主义、列宁主义、毛泽东思想,哲学,社会科学,自然科学,综合性图书五大基本部类。在此五大基本部类的基础上进一步划分为 22 个大类,分别以英文大写字母 A、B、C、D、E、F、G、H、I、J、K、N、O、P、Q、R、S、T、U、V、X、Z 区分。在基本大类表的基础上扩展而成基本类目表,以下是 R(医药、卫生)各级类目的简表:

R 医药、卫生

1 预防医学、卫生学

2 中国医学

3 基础医学

4 临床医学

5 内科学

6 外科学

71 妇产科学

72 儿科学

73 肿瘤学

74 神经病学与精神病学

75 皮肤病学与性病学

76 耳鼻咽喉科学

77 眼科学

78 口腔科学

79 外国民族医学

8 特种医学

9 药学

希望作者在投稿时根据文章内容选择标出分类号。